Published in final edited form as:

Int J Behav Dev. 2018 July ; 42(4): 445–452. doi:10.1177/0165025417743066.

Latent State-Trait Modeling: A New Tool to Refine Temperament Methodology

Jonathan Preszler and Maria A. Gartstein

Washington State University

Abstract

Questions concerning longitudinal stability and multi-method consistency are critical to temperament research. Latent State-Trait (LST) analyses address these directly, and were utilized in this study. Thus, our primary objective was to apply LST analyses in a temperament context, using longitudinal and multi-method data to determine the amount of trait vs. state variance, as well as convergence for measures of Distress to Limitations (DL) facets. Mothers' ratings and independent observations of DL behaviors collected on two occasions (8 months old and 12 months old) for 148 infants (49.2% female) were utilized. Single source latent state-trait (LST) analyses indicated that parent ratings of DL behavior (PDL) contained more trait ($M = 61\%$) than state residual ($M = 39\%$) variance, whereas independent observations (IO) of DL behavior contained substantially more state residual (75%) than trait (25%) variance. A multiple source LST analysis indicated virtually zero convergence for either trait or state residual variance between PDL and IO ratings ($M = 2\%$). In conclusion, PDL ratings were more trait-like across the four-month interval, whereas IO ratings of DL were more state-like in nature. Also, no convergence was found between the two methods of measurement. Results are discussed with an emphasis on implications for the utility of LST analyses in temperament research.

Keywords

temperament; distress to limitations; longitudinal research; latent state-trait

Introduction

One prominent controversy in temperament research concerns agreement among different sources of information, especially congruence between parent-report and observation-based indices of reactivity. Although laboratory observations are often characterized as more objective, parent-report indicators were shown to possess predictive validity that rivals alternative indicators (Pauli-Pott, Mertesacker, Bade, Haverkock, & Beckmann, 2003). In addition, it can be argued that maternal perspective on infant temperament is critical because of the primary caregiver's contribution to the child's milieu, especially in early childhood (Bornstein, 2014). At the same time, laboratory observations provide an opportunity to structure the situation in a standardized manner, administering episodes designed to elicit

certain attributes, such as distress to limitations or frustration (Gagne, Van Hulle, Aksan, Essex, & Goldsmith, 2011). More recently, it has been proposed that each methodological approach possesses unique strengths and weaknesses, and future research should articulate these distinctions more precisely (Gartstein et al., 2016). Convergence (or lack thereof) between parent-report and observation-based measures of temperament could be framed from the state vs. trait perspective. That is, the nature of parent-report and observation-based tools is such that the former lends itself primarily to capturing trait-related variability, whereas infant responses to brief, albeit highly structured experimental manipulations may speak more to state-related changes in emotion/reactivity.

Longitudinal data with multiple assessment points are required to answer state vs. trait questions, which are relevant not only to measurement considerations but also to temperament theories which typically include assumptions about temporal stability and setting/person consistency. Stability is viewed as one of the defining features of temperament, with significant consistency observed across assessments as early as infancy (Bornstein et al., 2015). At the same time, the psychobiological model of temperament anticipates changes as a function of development, especially rapid in early childhood, and considerable growth has been demonstrated for multiple attributes, such as fear (e.g., Gartstein et al., 2010). The current paper illustrates an application of a novel methodological approach, Latent State-Trait (LST) Analyses, to questions concerning stability/change and convergence of temperament measures. Specifically, examining the possibility that the commonly reported lack of convergence between laboratory observations and parent-report can be framed as a function of state vs. trait distinction among these indicators.

Existing studies addressing longitudinal stability and cross-setting consistency of temperament typically rely on correlation or regression coefficients (Bornstein et al., 2015; Garcia Coll et al., 1992), which do not directly address questions concerning stability and/or consistency of longitudinal variance. Studies relying on latent growth modeling (LGM) also fall short in this regard (e.g., Gartstein et al., 2010), as LGM addresses change in mean levels of an attribute, rather than variance stability. If the variance of an attribute (but not its mean values) lacks stability, the instability occurs in a non-systematic manner (some individuals decrease while others increase). Importantly, this pattern of results would suggest the attribute in question is relatively occasion-specific rather than trait-like.

LST analyses directly address stability and convergence questions using longitudinal/multi-method data. Single-method LST models apply to single-method longitudinal data (e.g., parent-report across multiple occasions) (for review of LST modeling, see Geiser, Hintz, Burns, & Servera, In press; Steyer, Mayer, Geiser, & Cole, 2014). With a minimum of two occasions of measurement, a single method LST model can separate the variance in temperament indicators into trait variance (i.e., see trait consistency definition in Table 1), state residual variance (i.e., see occasion-specificity definition Table 1), and measurement error variance (unreliability; Steyer et al., 2012). Trait variance represents the amount of true score variance in temperament indicators that is purely person-specific and independent of the occasion and/or person-occasion interactions. State residual (occasion-specific) variance, in contrast, represents the amount of true score variance in temperament indicators that reflects occasion-specific (situational) influences and/or person-occasion interactions. This

LST model can help determine if a temperament attribute is more trait- or more state-like for a single source across multiple occasions of measurement (Figure 1). Definitions for this model are presented in Table 1. A multiple method LST model (Courvoisier et al., 2008; Litson et al., 2016; Preszler Burns, Litson, Geiser, & Servera, 2016; Preszler et al., 2017) differs from a single method LST model by the simultaneous analysis of two or more methods of measurement (e.g., ratings). Including multiple methods in the same model (for example, information obtained from parent ratings and laboratory observations) makes it possible to study the convergent validity of trait and occasion-specific components across methods.

The multiple method LST model used in this study requires that one method be selected as a reference method (e.g., mother ratings) with the other methods considered nonreference methods (e.g., lab indicators). This multiple method LST model can be used to identify the amount of trait and state residual variance in temperament indicators that the nonreference methods either share or do not share with the reference method. Our multiple method LST model resulted in four variance components for the nonreference method (see Table 1)—(a) shared trait consistency, or the proportion of trait variance in temperament indicators that a nonreference method shares with the reference method; shared trait consistency thus indicates the degree of convergent validity with regard to stable (trait) aspects of behavior; (b) unique trait consistency, or the proportion of trait variance in temperament indicators that a nonreference method does not share with the reference method; unique trait consistency thus indicates the degree of method-specificity with regard to stable (trait) aspects of behavior; (c) shared occasion-specificity, or the proportion of occasion-specific variance in temperament indicators that a non-reference method shares with the reference method; shared occasion-specificity thus indicates the degree of convergent validity with regard to time-variable (state residual) aspects of behavior; and (d) unique occasion-specificity, or the proportion of occasion-specific variance in temperament indicators that is unique to a nonreference method and not shared with the reference method; thus, unique occasion-specificity indicates the degree of method-specificity with regard to variable (state residual) aspects of behavior. The decomposition of true score (i.e., reliable) variance into these four components for the nonreference method affords analysis of both temporal consistency and cross-method convergent validity of trait and state residual components in temperament indicators. Figure 2 shows the multiple method LST model used in the present study with the definitions for this model shown in the Table 1.

Importantly, there are different types of LST models that may be more appropriate depending on the specific goals of the investigation. For example, certain LST models can take into account autoregressive effects, and others can take into account trait change (see Geiser et al. [In press] for an in-depth explication of the variety of LST modeling techniques). The multitrait multistate (MTMS) model with no autoregressive or trait change effects (Geiser et al., 2017, figure 2) was selected for the purposes of this study, given that only two time points were utilized and temperament stability has been demonstrated within this developmental timeframe (e.g., Bornstein et al., 2015). This type of model was used for both the single- and multiple-method analyses, shown in our Figures 1 and 2, respectively.

LST models seem particularly well suited for the temperament context because these provide an opportunity to *directly* examine consistency of a temperament construct across time and across methods of measurement (e.g., lab measure and parent/observer ratings). Temperament is conceptualized as an enduring construct, yet empirical evaluations of stability remain informative, especially during periods of rapid development (e.g., infancy). Moreover, an ongoing measurement debate, elucidating agreement and areas of discordance among different sources of information, make this field optimal for LST applications.

Distress to Limitations as an example

Recent uses of LST involve investigations of developmental psychopathology, such as ADHD and ODD, in school-age children, (Litson, Geiser, Burns, & Servera, 2016; Preszler et al., 2016). However, to our knowledge, LST models have not yet been applied to temperament constructs, despite the demonstrated utility of this approach and its applicability to individual differences in reactivity and regulation. The present study was designed to address this gap in research and represents an illustration of LST in this context. For the purposes of this illustration, we focused on a fine-grained infant temperament construct of Distress to Limitations, defined as: “Fussing, crying or showing distress while a) in a confining place or position; b) in caretaking activities; c) unable to perform a desired action, and measured across two time points using two methods, reflecting anger/frustration. This level of analysis was undertaken because more global temperament factors (e.g., Negative Emotionality) represent multi-dimensional constructs with heterogeneous content. Distress to Limitations was also selected as a target because it has been successfully evaluated via structured laboratory observations, along with parent-report ratings (Stifter & Braungart, 1995; Gartstein & Rothbart, 2003). Although other fine-grained components of temperament should be examined via LST techniques in the future, we focused on Distress to Limitations for this initial illustration because negative emotionality has been the most widely studied broadband domain of temperament, and anger/frustration represents a component with clearly observable manifestations, expected to provide the basis for consistency across measurement modalities.

Our objectives were to (1) use single-method LST models to determine the amount of time-stable and occasion-specific variance for each method separately, and (2) to use a multiple-method model to determine the amount of shared time-stable variance and shared occasion-specific variance between the two methods: laboratory observation-based and parent-report scores. Stability was examined from 8 to 12 months, capturing an important developmental period at the end of the first year of life.

Method

Participants and Procedures

A community sample of 148 English-speaking mothers with healthy full-term infants was recruited through birth announcements and a universal prevention program. Participating caregivers (mean age = 28.67, SD = 5.27) were English speaking and primarily Caucasian (92%), representative of the general population. The vast majority of mothers were married (93.1%), having completed at least some college (90.9%), with family incomes generally >

\$20,000 (73.2%). Mothers responded to the Infant Behavior Questionnaire-Revised (Gartstein & Rothbart, 2003) at 8 and 12 months of age, and visited the laboratory with their infants (49.2% female; 50.8% male), who participated in the Laboratory Temperament Assessment Battery (Lab-TAB; Gagne, et al., 2011; Goldsmith & Rothbart, 1996). Only the Distress to Limitations (DL) scale of the IBQ-R and the Toy Retraction episode of Lab-TAB, designed to elicit mild frustration, were considered.

Measure

Infant Behavior Questionnaire-Revised (IBQ-R; Gartstein & Rothbart, 2003)—

The IBQ-R is a parent-report measure of infant temperament designed for infants between 3 and 12 months of age. The IBQ-R contains 191 questions regarding temperament characteristics, with frequency rated on a 7-point Likert scale (e.g., 1 = never, 5 = more than half the time, 7 = always). Questions regarding more common events concerned the past week, and less common events (e.g., meeting unfamiliar adults) the past 2-weeks. Distress to Limitations (DL) scale was utilized in this study, wherein its composition was scrutinized in order to meet the requirements of LST (see analyses and results sections for more information on item selection).

The Laboratory Temperament Assessment Battery (Lab-TAB; Goldsmith & Rothbart, 1996)—

The Laboratory Temperament Assessment Battery (Lab-TAB; Goldsmith & Rothbart, 1996), a widely used observation protocol, with multiple studies providing evidence of good psychometric properties (e.g., reliability; Diaz & Bell, 2012; Gagne, et al., 2011), was administered. The Toy Retraction episode was designed to elicit mild anger/frustration in the context of a situation encountered in daily life. Infants are seated in a high chair at a table, with the mother to his/her left and the experimenter to his/her right. The experimenter brings a toy into the infant's view, demonstrates how it works, and allows him/her to manipulate the object. After the infant has played with the toy for 15 seconds, the experimenter cues the mother to place the toy out of the infant's reach for 15 seconds. After 15 seconds, the toy is returned, and the procedure is repeated two more times, for a total of three 15-second trials.

Subsequently, research assistants were trained to assign reliable codes to infant behaviors reflecting anger/frustration in the Lab-TAB Toy Retraction task. These coders rated infant displays of anger/frustration across nine 5-second epochs for: (1) latency of first anger response; (2) intensity of anger expression; (3) intensity of distress vocalizations; (4) intensity of bodily anger; (5) presence of struggling/escape behavior; (6) intensity of struggling/escape behaviors; (7) attempts to reach toy; and (8) banging on the table (See Table 2 for the coding scheme description). Satisfactory reliability, including inter-rater agreement (75% to 88%), were demonstrated for these indicators (Majdandzic, van den Boom, & Heesbeen, 2008). In the current sample, inter-rater reliability coefficients ranged from .61 to .98 (mean $r = .81$).

Analyses

Measurement Requirements—Measurement Requirements must be met prior to model testing, including a well-fitting measurement model, invariance of loadings and intercepts/

thresholds, and mean stability across occasions. That is, similar to other applications of structural equation models (SEMs), LST models necessitate measurement models with adequate fit for each method of measurement (e.g., mother report and lab indicators). Furthermore, similar to other longitudinal and multiple group SEMs, LST requires invariance of loadings and intercepts/thresholds across time in the measurement models. Measurement invariance ensures that LST models are decomposing true score variance that is consistently interpreted by raters across time. Otherwise, decomposed variance components could be a function of a change in scale (or other factors) across time. Lastly, the LST model used in this study assumes mean stability, modeling time-stable and occasion-specific variance as independent of sample mean changes (i.e., trait changes). Thus, mean stability is tested in the measurement model before applying an LST techniques to the data.

In the present study, we followed a twofold strategy, first selecting DL items most relevant in content to the Toy Retraction experimental situation a-priori. Second, if measurement requirements (i.e., adequate measurement model, longitudinal invariance, and mean stability) were not satisfied for these a priori selected items, we sought indicators within the DL domain of the IBQ-R that met measurement requirements. Although in this study items were selected based on their content similarity to the Lab-TAB measure, as well as to each other, it should be noted that specific facets or subdomains within Distress to Limitations have not been previously identified.

Analyses were conducted using Mplus 7.4 (Muthén & Muthén, 2012). The robust maximum likelihood estimation (MLR) estimator accounted for any non-normality in the item ratings and Full Information Maximum Likelihood (FIML) was used to accommodate missing data. The comparative fit index (CFI, ideal study criterion .95), the Tucker-Lewis Index (TLI, ideal study criterion .95), and the root mean square error of approximation (RMSEA, ideal study criterion .05) were considered to evaluate model fit. Before applying the LST model, we first needed to demonstrate invariance of like-item loadings and intercepts across time. First, invariance analyses were conducted to ensure equivalence of like-item loadings and intercepts across time using changes in CFI value and chi-square difference tests. If the CFI decrease was $< .01$ with the introduction of constraints on like-item loadings and like-item thresholds (strong invariance), the invariance constraints were considered tenable (Little, 2013, Chapter 5). In addition, if the chi-square difference test was non-significant at $p < .05$, we considered the invariance constraints tenable.

Results

As noted, we initially focused on Distress to Limitations items that overlapped in content with the experimental situation – i.e., addressed anger/frustration responses to toy removal (i.e., “When something the baby was playing with had to be removed, how often did s/he: (1) cry or show distress for a time; (2) seem not bothered?; and (3) “When the baby wanted something, how often did s/he become upset when s/he could not get what s/he wanted?”). However, these did not produce an adequate-fitting configural model across time (RMSEA = .092; CFI = .966; TLI = .914), and the third item did not have high loadings (.54) on the construct at either time point.

Additional parent-report DL (PDL) items with content addressing two distinct themes were considered next. The first item set included three questions related to waking up (e.g., “After sleeping, how often did the baby (1) fuss or cry immediately?”, (2) “play quietly in crib?”, and (3) “cry if someone doesn’t come within a few minutes?,,). The second set included three items related to behavior before falling asleep (e.g., “How often did the baby: (1) seem angry (crying and fussing) when you left her/him in the crib? (2) seem contented when left in the crib? (3) cry or fuss before going to sleep for naps? Only the first of these item sets met invariance (CFI = .000, χ^2 difftest = 2.25, $p = .691$) and relative mean stability assumptions necessary to conduct our LST analyses. Thus, DL items addressing anger/frustration upon waking up were considered further. A laboratory-based measurement model comprised of three independent observer (IO) indices met strong measurement invariance (CFI = .008, χ^2 difftest = 7.25, $p = .123$) and mean stability requirements of LST analysis: bodily anger, intensity of anger, and distress vocalizations.

Time invariance of loadings and thresholds was attained across both methods. We conducted invariance analyses of like-item loadings and intercepts across time on the PDL and IO measures. The configural model provided a good fit for each method (CFIs > .997, TLIs > .989, and RMSEAs < .035). Invariance of like-item loadings and intercepts held across the two time points for each of the models (i.e., no decrease in CFI was greater than .008), providing a foundation for proceeding with LST analyses.

PDL and IO LST models were associated with good fit (Table 3), as both estimated models adequately represented the data. The multiple-method model, including both methods simultaneously, had good fit (Table 3). The variance components of our models are displayed in Table 4. PDL single method analyses showed more trait consistency ($M = .61$) than occasion-specificity ($M = .39$), whereas the IO single method results demonstrated substantially more occasion-specificity ($M = .75$) than trait consistency ($M = .25$). The multiple-method model, comparing PDL to the IO, showed the IO measure shared virtually no variance with PDL, as indicated by the levels of shared trait consistency ($M = 0\%$) and shared minimal occasion-specificity ($M = 2\%$), in contrast to the higher levels of unique trait consistency ($M = 28\%$) and unique occasion-specificity ($M = 70\%$).

Discussion

Our study sought to introduce the use of LST analysis (Steyer et al., 2014) to investigate the trait vs. state components of temperament assessed via parent-report and laboratory observations. The longstanding questions concerning agreement among different sources of information and stability vs. change in individual differences make this approach ideal for temperament research. Our initial effort focused on Distress to Limitations, reflecting anger and frustration-related emotional reactivity exhibited across contexts, including a situation wherein an attractive toy is removed, enacted in the laboratory evaluation of individual differences.

Single-method analyses indicated that PDL was about 60% stable across two time points, but IO was only about 25% stable. Although these findings require replication, this pattern of results suggests a number of possibilities. First, these findings bring attention to the

questions concerning test-retest reliability of IO ratings of infants; specifically, those concerning an appropriate interval that would capture this psychometric property of the test, rather than the development/instability of the phenomenon in question. Although short in duration, the 8 to 12 period could represent a transition with respect to temperament, and development more generally (Bornstein, Arterberry, & Lamb, 2014), and it may be more appropriate to frame IO as more sensitive to these developmental shifts. Laboratory-based measures were previously described as advantageous in terms of sensitivity to developmental changes, as parents tend to be stable in their expectations, not shifting perceptions of child attributes effectively with developmental transitions (Gagne et al., 2011; Saudino, 2003). This increased sensitivity is especially critical during periods of rapid transitions, and observation-based techniques may be preferred to parent-report in this context.

According to the multiple method analysis, the two methods (parent-report and laboratory observations) shared virtually zero variance. There are many potential explanations, including the significantly briefer time frame of the independent observation compared to maternal ratings, the introduction of an unfamiliar person into the infant's environment in the context of Lab-TAB, or other novel aspects of the laboratory. Moreover, discrepancy in content of questionnaire items and the nature of laboratory observations could be responsible, as the most content-consistent items did not satisfy LST requirements. That is, although the parent-report IBQ-R indicators and Lab-TAB derived codes reflect the same underlying temperament construct, namely Distress to Limitations, these did not overlap in terms of specific content. That is, items derived from laboratory observations addressed toy removal, whereas parents reported concerning anger/frustration upon waking up. Distress to Limitations was conceptualized (Rothbart, 1981) and has been consistently treated as a unidimensional construct (Gartstein & Rothbart, 2003; Gartstein, Bridgett, & Low, 2012). The pattern of results observed in this study suggests this approach should be subjected to a systematic empirical evaluation. Notably, these latter explanations speak to a context-dependent nature of DL in infancy. Although replication and future research utilizing additional measures (e.g., physiological indicators) is required, our findings suggest that parent-report of infant DL functions more consistently with the trait-like definition than IO ratings recorded in the laboratory setting. Laboratory-based observations may appear as more state-like in their functioning due to increased sensitivity to developmental changes occurring in the 2nd half of the first year of life (e.g., Gartstein, Hancock, & Iverson, in press).

Limitations

The present sample was demographically homogeneous, limiting generalizability of results. Only one temperament dimension, Distress to Limitations was examined, with observations limited to two time points. In addition, we relied on heterogeneous aspects of the "Distress to Limitations" construct across the two examined methods, as a result of the necessarily limited content addressed in the laboratory, and the failure of parallel parent-report items to meet measurement requirements. Unfortunately, results of this study do not provide the basis for conclusive interpretations concerning this failure of items most comparable with respect to face validity to meet LST prerequisite criteria. However, if replicated, this pattern of

results would have implications for temperament measurement and theory. It may be that manifestations of certain temperament attributes are context dependent during specific developmental periods, such as infancy. That is, whereas our laboratory observation-based indicators reflected infant distress in response to toy removal, parental ratings addressed child distress in response to waking-up. This discrepancy stands in contrast to recent studies of ADHD (Litson et al., 2016) and ODD (Preszler et al., 2016), wherein different raters completed nearly identically worded items. The latter technique appears useful in minimizing item-specific method effects that might otherwise lead to higher nonconvergence across sources, and could be applied in the context of temperament research.

Future Directions

Future studies should address the limitations noted above, recruiting more diverse samples with respect to socio-demographic factors. Additional time points should also be collected to help elucidate causes of instability across time and/or inconsistency across settings. More extensive longitudinal designs, following participants to determine if the trait-like presentation becomes consolidated later in childhood, should also be implemented. Questions concerning dimensionality of temperament constructs should also be addressed more conclusively in future research. For example, facets addressing components of this attribute (e.g., frustration to blocked goals) could be examined to determine if these provide meaningful distinctions. Importantly, researchers should continue to develop and refine tools to support multi-method measurement of different aspects of temperament, ensuring good psychometric properties. Methods employed in related LST applications, for example Burns & colleagues' studies (Litson et al., 2016; Preszler et al., 2016, 2017) addressing adolescent and child psychopathology (e.g., ADHD, ODD, SCT), could provide a useful model in their rigorous psychometric evaluations preceding LST studies (Burns & Lee, 2011; Burns, Servera, Bernad, Carillo, & Severa, 2014; Kadka & Burns, 2013).

Conclusion

Temperament is typically conceptualized as trait-like, expected to manifest consistently across time and settings. We set out to demonstrate that LST modeling allows for a more direct and effective investigation of this trait-like nature, relative to previously used correlational, regression, and LGM techniques. Our illustration demonstrated time-stability for mothers' DL ratings, but little time-stability in IO indicators, or consistency between the two sets of measures designed to address the same underlying temperament construct.

Acknowledgments

Jonathan Preszler's work was supported by the Anthony Marchionne Foundation for the Scientific Study of Human Relations and Psychological Processes Endowed Graduate Fellowship for Research at Washington State University.

References

- Bornstein MH. Parents' reports about their children's lives. In: Ben-Arieh A, Cashmore J, Goodman GS, Melton GB, editors *The SAGE handbook of child research*. London, UK: Sage; 2014. 486–533.
- Bornstein MH, Arterberry ME, Lamb ME. *Development in infancy: A contemporary introduction*. 5th. New York: Psychology Press; 2014.

- Bornstein MH, Putnick DL, Gartstein MA, Hahn C, Kwak K, Auestad N, O'Connor DL. Infant Temperament across the First Years of Postnatal Life: Stability by Age, Gender, Birth Order, Term Status, and SES. *Child Development*. 2015; 86:844–863. [PubMed: 25865034]
- Burns GL, Lee S. Child and adolescent disruptive behavior inventory–Parent version 5.0. Pullman, WA: Author; 2011.
- Burns GL, Servera M, Bernad M, Carrillo JM, Geiser C. Ratings of ADHD symptoms and academic impairment by mothers, fathers, teachers and aides: Construct validity within and across settings as well as occasions. *Psychological Assessment*. 2014; 26:1247–1258. DOI: 10.1037/pas0000008 [PubMed: 24932644]
- Diaz A, Bell MA. Frontal EEG asymmetry and fear reactivity in different contexts at 10 months. *Developmental Psychobiology*. 2012; 54:536–545. DOI: 10.1002/dev.20612 [PubMed: 22006522]
- Gagne JR, Van Hulle CA, Aksan N, Essex MJ, Goldsmith HH. Deriving childhood temperament measures from emotion-eliciting behavioral episodes: Scale construction and initial validation. *Psychological Assessment*. 2011; 23:337–353. DOI: 10.1037/a0021746 [PubMed: 21480723]
- Garcia Coll CT, Halpern LF, Vohr BR, Seifer R, Oh W. Stability and correlates of change of early temperament in preterm and full-term infants. *Infant Behavior & Development*. 1992; 15:137–153. DOI: 10.1111/j.1467-7687.2010.00969.x
- Gartstein MA, Bridgett DJ, Rothbart MK, Robertson C, Iddins E, Ramsay K, Schlect S. A Latent Growth Examination of Fear Development in Infancy: Contributions of Maternal Depression and the Risk for Toddler Anxiety. *Developmental Psychology*. 2010; 46:651–668. [PubMed: 20438177]
- Gartstein MA, Putnam SP, Aaron E, Rothbart MK. Temperament and personality. In: Matzman S, editor *Oxford Handbook of Treatment Processes and Outcomes in Counseling Psychology*. New York, NY: Oxford University Press; 2016. 11–41.
- Gartstein MA, Rothbart MK. Studying infant temperament via the revised infant behavior questionnaire. *Infant Behavior and Development*. 2003; 26:64–86. DOI: 10.1016/S0163-6383(02)00169-8
- Geiser C, Hintz F, Burns GL, Servera M. Latent variable modeling of personsituation data. In: Funder D, Rauthmann J, Sherman R, editors *Handbook of psychological situations*. Oxford Press; (in press)
- Goldsmith HH, Rothbart MK. The Laboratory Temperament Assessment Battery. Prelocomotor Version, V.3.00. 1996. Unpublished manuscript
- Khadka G, Burns GL. A measurement framework to determine the construct validity of ADHD/ODD rating scales: Additional evaluations of the Child and Adolescent Disruptive Behavior Inventory. *Journal of Psychopathology and Behavioral Assessment*. 2013; 35:283–292. DOI: 10.1007/s10862-012-9336-z
- Litson K, Geiser C, Burns GL, Servera M. Trait and state variance in multi-informant assessments of ADHD and academic impairment in Spanish first-grade children. *Journal of Clinical Child & Adolescent Psychology*. 2016 Feb.4416:1–14. DOI: 10.1080/15374416.2015.1118693
- Little TD. Longitudinal structural equation modeling. New York: Guilford; 2013.
- Majdandzic M, van den Boom DC, Heesbeen DG. Peas in a pod: biases in the measurement of sibling temperament? *Developmental Psychology*. 2008; 44:1354–1368. DOI: 10.1037/a0013064 [PubMed: 18793068]
- Muthén LK, Muthén BO. *Mplus User's Guide*. Sixth. Los Angeles, CA: Muthén & Muthén; 1998–2010.
- Pauli-Pott U, Mertesacker B, Bade U, Haverkock A, Beckmann D. Parental perceptions and infant temperament development. *Infant Behavior & Development*. 2003; 26:27–48. DOI: 10.1016/S0163-6383(02)00167-4
- Preszler J, Burns GL, Litson K, Geiser C, Servera M. Trait and state variance in oppositional defiant disorder symptoms: A multi-source investigation with Spanish children. *Psychological Assessment*. 2016; 28(4):1–13.
- Preszler J, Burns GL, Litson K, Geiser C, Servera M, Becker SP. How consistent is sluggish cognitive tempo across occasions, sources, and settings? Evidence from latent state-trait modeling. *Assessment*. 2017.

- Saudino KJ. Parent ratings of infant temperament lessons from twin studies. *Infant Behavior & Development*. 2003; 26:100–107. DOI: 10.1016/S0163-6383(02)00171-6
- Steyer R, Mayer A, Geiser C, Cole D. A Theory of States and Traits — Revised. *Annual Review of Clinical Psychology*. 2014 Jul.10:1–28. DOI: 10.1146/annurev-clinpsy-032813-153719

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 1.

Single method latent state-trait model with parcel-specific trait factors measured at two time points. This model was applied to both PDL and IO methods separately. The Mplus code for this model is shown in the supplemental materials. T = trait factor; O = state (occasion-specific) residual factor. Participant $n = 147$.

Figure 2.

Multiple method latent state-trait model with item-specific trait factors. T = reference trait factors that are defined by mother reports (Shared Trait Consistency); TM = residual trait factors that pertain to the lab indicators (Unique Trait Consistency); O = reference occasion-specific factors that are defined by mother reports (Shared Occasion-Specificity); OM = occasion-specific factors that pertain lab indicators (Unique Occasion-Specificity). M_{11} and M_{21} = IBQ19; M_{12} and M_{22} = IBQ20R; M_{13} and M_{23} = IBQ21; L_{11} and L_{21} = BDYANG; L_{12} and L_{22} = INANG; L_{13} and L_{23} = INDSVC. Participant $n = 147$.

Table 1

Definitions for Single and Multiple Source Latent State Trait (LST) Models

Term	Definition
Reliability	The proportion of observed score variance that is due to systematic sources of variance rather than measurement error.
<i>Single Method LST Variance Definitions</i>	
Trait Consistency	The proportion of reliable variance that reflects stable trait variance. Consistency reflects reliable variance that is purely person-specific and independent of the situation and/or person-situation interaction.
Occasion Specificity	The proportion of reliable variance that reflects occasion-specific influences and/or person-situation interactions.
<i>Multiple Method LST Variance Definitions</i>	
Shared Trait Consistency	The proportion of trait variance in a non-reference method (i.e., lab ratings) that is <i>shared</i> with the reference method (i.e., mother ratings). Indicates convergent validity for trait components.
Unique Trait Consistency	The proportion of trait variance in a non-reference method (i.e., lab ratings) that is <i>not shared</i> with the reference method (i.e., mother ratings). Indicates method-specificity of trait components.
Shared Occasion-Specificity	The proportion of occasion-specific variance in a non-reference method (i.e., lab ratings) that is <i>shared</i> with the reference method (i.e., mother ratings). Indicates convergent validity of occasion-specific components.
Unique Occasion-Specificity	The proportion of occasion-specific variance in a non-reference method (i.e., lab ratings) that is <i>not shared</i> with the reference method (i.e., mother ratings). Indicates method-specificity of occasion-specific components.

Note. “Reliable variance” or “reliability” is also sometimes referred to as “true score variance.” Similarly, “occasion-specific variance” is also sometimes referred to as “state residual variance” in LST modeling.

Table 2

Toy retraction coding scheme

Variable	Coding Scheme: Rating Scales
1. Latency of first anger response	Time, in seconds, starting when mother removes toys from child to the first sign of anger (facial, vocalic, postural, or instrumental; everything but approach). The first definite anger response is any response that would be coded as a 1 or higher below.
2. Intensity of anger expression	0 = No facial region shows codable movement. 1 = Only one facial region shows codable movement, identifying a low intensity anger, or expression is ambiguous. 2 = Only 2 facial regions show codable movement, or expression in one region is definite. 3 = An appearance change occurs in all 3 facial regions, or coder otherwise has impression of strong anger.
3. Intensity of distress vocalizations	0 = No distress. 1 = Mild vocalization, may be difficult to identify as hedonically negative. 2 = Definite whimpering, limited to a short (1–2 seconds) duration. 3 = Longer whining, fussing, mild protest, or low-intensity cry (cry has extended or rhythmic quality). 4 = Definite non-muted crying. 5 = Full intensity cry/scream.
4. Intensity of bodily fear	0 = No sign of bodily fear. 1 = Decreased activity: apparent and sudden decrease in activity level of the child. 2 = Tensing: visible tensing of the muscles and decreased activity. 3 = Freezing or trembling: tensing of the entire body with no motion, or trembling due to extreme muscular tension.
5. Presence of struggling	Presence or absence of escape behavior (e.g., when the child leans to the side of the chair, or forward, and/or attempts to push him/herself out of the chair, using his/her arms or legs).
6. Intensity of struggling	0 = No struggling to hold on to the toy at all. No resistant movement. 1 = Low intensity struggling. 2 = Medium intensity struggle. Sporadically pulls/attempts to get out of the chair. Movements could include medium intensity pulling, pushing up (using hands or legs), leaning, arching back, or kicking. Generally lasts 2–3 seconds. 3 = Moderately high intensity struggle. Near-continuous moderate-intensity pulling or leaning. Can include the same movements as number 2 with higher intensity. Generally lasts 3 – 4 seconds. 4 = High intensity struggle. Continuous movement of moderately high intensity with intervals of high intensity movements to get out of the chair. Generally lasts five seconds.
7. Attempts to reach the toy	Presence or absence of attempts to reach toward the toy is recorded.
8. Presence of bodily anger	Presence or absence of detectable bodily anger (e.g., bodily tensing)
9. Banging on table	Presence or absence of banging on the table is recorded for each epoch.

Table 3
Goodness-of-Fit Statistics for Latent State Trait Analyses for Distress to Limitations Reports

Model	df	χ^2	CFI	TLI	RMSEA (90% CI)
Single Method Latent State Trait Analyses					
Parent-rated Distress to Limitations	8	11.16 <i>ns</i>	.987	.975	.052 (.000 – .118)
Independent Observer-rated Distress to Limitations	11	16.72 <i>ns</i>	.986	.981	.059 (.000 – .114)
Multiple Method Latent State Trait Analysis					
Multiple Method	48	64.09 <i>ns</i>	.980	.973	.048 (.000 – .076)

Note. CFI = comparative fit index; TLI = Tucker-Lewis index; RMSEA = root mean square error of approximation; CI = confidence interval

ns = nonsignificant.

* = $p < .05$. Participant $n = 147$

Table 4

Average Consistency, Occasion-Specificity, and Reliability Estimates from Single and Multiple Method Latent State-Trait Analyses on the Distress to Limitations Reports

Single Method				
Method		Trait Consistency	Occasion-Specificity	Reliability
Parent-rated Distress to Limitations		.61 (.40, .91)	.39 (.09, .60)	.79 (.65, .98)
Independent Observer-rated Distress to Limitations		.25 (.07, .44)	.75 (.56, .93)	.86 (.76, .98)

Multiple Method				
Method	Shared Trait Consistency	Unique Trait Consistency	Shared Occasion-Specificity	Unique Occasion-Specificity
Multiple Method	.00 (.00, .00)	.28 (.15, .44)	.02 (.00, .06)	.70 (.52, .84)

Note. Entries indicate averages across items (with ranges in parentheses). Multiple Method model compares Independent Observer-rated factor to Parent-rated factor. Participant $n = 147$.