Title

The Structure of Adolescent Temperament and Associations With Psychological Functioning: A Replication and Extension of Snyder et al. (2015)

Journal

Journal of Personality and Social Psychology, 121(5)

ISSN

0022-3514

Authors

Lawson, Katherine M
Atherton, Olivia E
Robins, Richard W

Publication Date

2021-11-01

DOI

10.1037/pspp0000380

Peer reviewed

HHS Public Access

Author manuscript

J Pers Soc Psychol. Author manuscript; available in PMC 2022 November 01.

The Structure of Adolescent Temperament and Associations with Psychological Functioning: A Replication and Extension of Snyder et al. (2015)

Katherine M. Lawson[1], **Olivia E. Atherton**[1], **Richard W. Robins**[1]

[1]University of California, Davis

Abstract

The present study attempts to replicate and extend Snyder et al. (2015, *JPSP*). The original study examined the latent factor structure of the EATQ-R, a commonly used measure of adolescent temperament, and showed that the resulting latent factors (i.e., Effortful Control, Negative Emotionality, and Positive Emotionality) had theoretically meaningful concurrent associations with several measures of adolescent functioning (depression, anxiety, ADHD, relational aggression, and school performance and behavior). We performed these same analyses using data from a large sample of Mexican-origin youth (*N*=674), and also examined prospective associations between the three EATQ-R factors and measures of adolescent functioning assessed two years later. We found some evidence supporting the bifactor models reported in the original study but poor replication of the correlations among latent factors. Additionally, model comparisons demonstrated that correlated factors models led to more interpretable factors than the bifactor models. In contrast, we replicated most of the concurrent correlations (and extended the findings to prospective associations) between the EATQ-R factors and measures of adolescent functioning, supporting the construct validity of the EATQ-R as a measure of adolescent temperament. Thus, these findings raise concerns about the generalizability of the factor structure identified by Snyder et al. (2015), but bolster claims about the generalizability of the concurrent and predictive validity of the EATQ-R. Overall, differences between the present findings and those of Snyder et al. highlight the importance of ongoing construct validation in youth temperament research, especially with participants from groups traditionally underrepresented in psychological research.

Keywords

conceptual replication; generalizability; temperament; adolescence

Introduction

The path to a cumulative, replicable science of personality requires the use of psychometrically sound measures (Flake, Pek, & Hehman, 2017). One area of personality psychology that is underdeveloped in terms of construct validation is youth temperament

Corresponding author: Katherine M. Lawson, M.A. Department of Psychology University of California, Davis One Shields Ave. Davis, CA 95616, kmlawson@ucdavis.edu.

research, which examines individual differences in reactivity and self-regulation that are present from an early age and relatively enduring (Rothbart, 2007; Rothbart, 2011). There is no agreed upon taxonomy of the most important dimensions of youth temperament or consensus about which measures to use (Tackett & Durbin, 2017). This chaotic state impedes progress toward understanding how temperament differences develop during the highly consequential adolescent years. Recently, Tackett and Durbin (2017) warned the field that youth temperament research is "in dire need of more sophisticated and thoughtful measurement work", and proposed that researchers should "contribute to the long-standing but tedious journey to construct validity" (p. 1).

The present study heeds Tackett and Durbin's call to action and focuses on the construct validity of the most widely used measure of adolescent temperament - the *Early Adolescent Temperament Questionnaire – Revised* (EATQ-R; Ellis & Rothbart, 2001).[1] The EATQ-R was designed as a self- and informant report of temperament for youth aged from 9- to 15-years, but it is commonly used with older participants, including those up to 19-years old (e.g., Snyder et al., 2015).[2] The questionnaire assesses the three key constructs in Rothbart's highly influential temperament model (Rothbart, Ahadi, & Evans, 2000): effortful control (EC), negative emotionality (NE), and positive emotionality (PE). According to Rothbart's model, the EC domain involves one's capacity to plan and suppress inappropriate impulses (Inhibitory Control), perform an action or pursue goals when there are competing desires (Activation Control), and focus and shift attention when needed (Attention). The NE domain involves unpleasant affect derived from anticipating distress (Fear), negative affect related to ongoing tasks being interrupted (Frustration), and behavioral inhibition to social interaction (Shyness). The PE domain involves pleasure derived from high intensity or novel activities (High Intensity Pleasure/Surgency; hereafter referred to as Surgency), pleasure derived from low-intensity environmental stimulation (Pleasure Sensitivity), awareness of low-intensity environmental stimulation (Perceptual Sensitivity), and a desire for close, warm interpersonal connections (Affiliation). The EATQ-R also includes two additional scales: Aggression (hostile reactivity to negative stimuli including person- and object-directed violence) and Depressed Mood (unpleasant affect, lowered mood, and lack of enjoyment in activities). These scales were not considered part of any temperament domain in the original development of the EATQ-R, but they are conceptually linked to the NE domain, and are sometimes scored as part of the NE superordinate factor, along with the Fear, Frustration, and Shyness scales.

The development of the EATQ-R was guided by Rothbart and colleague's broad theory of temperament (e.g., Rothbart & Ahadi, 1994; Rothbart & Bates, 2006). This theory suggests that the structure of temperament is captured by reactivity (NE, PE) and regulation (EC), and specifies how each domain (as measured by the EATQ-R, or any other Rothbart-developed temperament questionnaire[3]) should be associated with real-world outcomes. In particular,

[1]The article describing the initial development of the EATQ (Capaldi & Rothbart, 1992) has been cited almost 500 times, and the publication describing the development of the revised EATQ-R (Ellis & Rothbart, 2001) has been cited over 600 times.
[2]More information about the EATQ-R is available at the following website: https://research.bowdoin.edu/rothbart-temperament-questionnaires/instrument-descriptions/the-early-adolescent-temperament-questionnaire/
[3]Rothbart and colleagues developed a suite of age-appropriate temperament measures, including measures designed for infancy, early childhood, childhood, middle childhood, adolescence (i.e., the EATQ-R), and adulthood.

internalizing problems (e.g., Anxiety, Depression) should be associated with high levels of NE (Brandes et al., in press; Nigg, 2006), whereas externalizing problems (e.g., ADHD) should be associated with high levels of NE and PE and low levels of EC (Brandes et al., in press; Eisenberg et al., 2007; Eisenberg et al., 2009; Muris, Meesters, & Blijlevens, 2007). Additionally, worse interpersonal functioning should be associated with low levels of EC and high levels of NE (Coplan & Bullock, 2012; De Bolle & Tackett, 2013; Eisenberg et al., 2009; Ojanen, Findley, & Fuller, 2012; Tackett et al., 2014). Further, better academic performance and school behavior should be associated with high levels of EC and low levels of NE (Valiente et al., 2013), but not associated with PE. Assuming Rothbart's theory is valid, the expected theoretical associations between the temperament domains and real-world outcomes allow researchers to evaluate the construct validity of her measure of temperament, the EATQ-R. Specifically, concurrent relations between temperament domains and life outcomes provide an opportunity to assess the extent to which there is *concurrent validity* of the EATQ-R, whereas temperament domains predicting future life outcomes assesses *predictive validity* of the measure. Past work on the concurrent and predictive validity of the EATQ-R has yielded mixed findings, in part because the field lacks consensus about how best to score the EATQ-R to assess the presumed underlying constructs (i.e., EC, NE, PE) (see Footnote 2 in Snyder et al., 2015).

The mixed findings in past research may also be due to problems with the internal structure of the measure. Notably, Rothbart's temperament measures were developed through a theory-driven, top-down approach with very little empirical, bottom-up validation (e.g., Kim, Brody, & Murry, 2003; Muris & Meesters, 2009). The limited work examining item-level analyses of Rothbart's temperament measures suggests that there are structural problems with many of Rothbart's temperament measures, including the EATQ-R. Kotelnikova and colleagues (2016) examined the structure of the Children's Behavior Questionnaire (CBQ; Rothbart, Ahadi, & Hershey, 1994) and found little evidence for the theorized higher-order three-factor structure and many items did not load well onto their designated lower-order factor. Kotelnikova and colleagues (2017) examined these same properties in the Temperament in Middle Childhood Questionnaire (TMCQ; Simonds & Rothbart, 2004) and found similar problems with both the higher-order and the lower-order structure. Other than Snyder et al. (2015), only one study has examined the factor structure of the EATQ-R. In particular, Latham et al. (2020), using data from a large sample of Australian adolescents (10–12 years old), found poor fit for a hierarchical model that specified four higher-order factors (EC, NE, Surgency, Affiliativeness) along with subfactors corresponding to the EATQ-R subscales. However, they found that fit improved substantially when items were allowed to have cross-loadings on subfactors other than their designated subfactor (e.g., an Inhibitory Control item was allowed to load onto both the Inhibitory Control and the Activation Control subfactors), and when subfactors were allowed to have cross-loadings on domains other than their designated domain (e.g., Inhibitory Control was allowed to load onto both the EC and Affiliativeness domains). Together, these findings suggest that Rothbart's temperament measures have poor structural validity. However, despite these psychometric issues, there is still substantial evidence for the convergent and discriminant validity of the EATQ-R (Muris & Meesters, 2009), which is consistent with

the notion that internal structure is only one element of construct validity (Hopwood & Donnellan, 2010; Loevinger, 1957).

Snyder et al. (2015)

In 2015, Snyder and colleagues published an article in the *Journal of Personality and Social Psychology* (*JPSP*) that endeavored to clarify the latent structure of the EATQ-R and the extent to which the underlying factors showed theoretically expected associations with various measures of adolescent functioning. Specifically, the authors used data from a large sample of adolescents (N = 2,026) to identify the factor structure of the EATQ-R via confirmatory factor analysis (CFA) and then examined associations between the resulting factors and concurrent measures of depression, anxiety, ADHD, peer interactions, and school grades and behavior. That is, Snyder and colleagues examined the extent to which the latent factors of the EATQ-R demonstrate concurrent validity with theoretically related life outcomes.

Snyder et al. (2015) used bifactor CFA models to assess the structure of the EATQ-R. These models allow researchers to study how items on a scale are related to both the general and domain-specific aspects of a broader latent construct (Chen, West, & Sousa, 2006; Reise, Moore, & Haviland, 2010) and are increasingly popular in research on personality and psychopathology (e.g., when studying the *p* factor; Caspi et al., 2014). In these models, each item can load onto a common factor, a domain-specific factor, or both a common and a domain-specific factor. These item loadings provide information about how to conceptualize the hierarchical structure of the construct. Further, the common and domain-specific factor scores can be related to external variables to better understand the psychological meaning of the shared (i.e., common) and unique (i.e., domain-specific) aspects of the construct.

Results from the best-fitting bifactor models in Snyder et al. (2015) generally support Rothbart's theoretical conception of the EC and NE domains. In particular, most of the variance of EC was captured by a common general factor, with the remaining variance being accounted for by a specific Activation Control factor. For NE, the model that best fit the data included a common general NE factor and several specific factors corresponding to each of the NE subscales (i.e., Fear, Frustration, Shyness, Aggression, Depressed Mood). However, the latent structure of PE was more complicated than expected. In particular, there was no psychologically coherent common factor for PE; instead, the Surgency items, which are arguably the theoretical core of the PE construct, formed a completely separate factor that was not strongly correlated (r = .29) with a common factor comprised of the remaining PE scales (Pleasure Sensitivity, Perceptual Sensitivity, Affiliation). The disjointed structure of PE led Snyder and colleagues to conclude that the PE subscale may actually assess two distinct constructs (i.e., surgency/sensation-seeking and general sensitivity to stimuli) rather than a single coherent domain of positive emotionality.

After examining the latent factor structure of the EATQ-R via bifactor models, Snyder et al. (2015) investigated the extent to which the resulting common and specific factors show concurrent associations with various indicators of adolescent functioning. Consistent with Rothbart's theoretical model and past empirical work, high EC was associated with decreased psychopathology (i.e., depression, anxiety, ADHD), decreased antisocial behavior

toward and victimization by peers, higher grades, and fewer school discipline problems. Also consistent with theory and past work, the common NE factor was associated with more instances of relational aggression and increased psychopathology, with specific NE factors differentially predicting specific psychopathology symptoms. In particular, and not surprisingly, the specific Depressed Mood factor was associated with depression symptoms, the specific Fear factor was associated with Harm Avoidance symptoms of anxiety, and the specific Aggression factor was associated with more relational aggression, whereas there were no significant concurrent associations between measures of adolescent functioning and the Frustration-specific and Shyness-specific factors. Unlike EC and NE, the PE factors had more complex associations with adolescent functioning; for example, the Surgency factor was only associated with fewer separation/panic symptoms of Anxiety, whereas the common PE factor was positively associated with harm avoidance symptoms of Anxiety. Additionally, the specific Affiliation factor was positively associated with Anxiety and Depression symptoms as well as both antisocial behavior toward peers and victimization by peers. Further, the common PE factor was positively associated with relational aggression (both perpetration and victimization). Together, this pattern led Snyder and colleagues to conclude that PE as measured by the EATQ-R may be tapping into a broad sensitivity to one's environment, rather than positive emotionality as conceptualized by Rothbart and colleagues.

The work of Snyder et al. (2015) provides an important and timely empirical investigation of the latent structure of adolescent temperament as measured by the EATQ-R, as well as the extent to which the EATQ-R has concurrent validity with respect to relevant real-world outcomes. Moreover, the authors present results from their bifactor CFAs as an updated, empirically-derived scoring method for use in future research aiming to measure adolescent temperament via the EATQ-R. However, there are several limitations to Snyder et al.'s work, which we discuss below.

Methodological Limitations—Snyder et al. (2015) concluded that a bifactor model fits their EATQ-R data best. However, there is evidence of statistical bias in favor of bifactor models, even when another (simpler) model fits similarly well (Murray & Johnson, 2013). Thus, when comparing bifactor models with other models, it is especially important to consider multiple model fit indices that take model parsimony into account (Murray & Johnson, 2013), which Snyder et al. did not do. In addition, there is evidence that bifactor models are vulnerable to overfitting, meaning that "model fit statistics are unreliable indicators of the validity of bifactor models" (Watts, Poore, & Waldman, 2019, p. 3) and the resulting parameter estimates are often unstable and/or difficult to interpret. Together, these two issues suggest that model fit indices may have trouble differentiating between bifactor and other models, and so substantive issues should also be considered when determining which factor model to retain (Morgan, Hodge, Wells, & Watkins, 2015).[4]

These methodological issues highlight potential problems with Snyder et al.'s bifactor model findings. In particular, they used change in chi-square to compare models, which is notoriously sensitive to large sample sizes (Cheung & Rensvold, 2002), and only use a

[4]Notably, many of these concerns about bifactor models were not widely discussed until after Snyder et al. conducted their research.

single fit index (RMSEA) that takes into account model parsimony (Hu & Bentler, 1998). Importantly, Snyder et al. found that the correlated factors model fit almost as well as the bifactor model (see Table 1, Snyder et al., 2015), which makes their selection of the bifactor model somewhat problematic given what we now know about biased fit indices and problems of overfitting with these models. The present study attempts to address some of these issues by comparing several different factor structures (i.e., a single-factor model, a correlated-factors model, a modified bifactor model, and a hierarchical model) using multiple fit indices (i.e., chi-square, RMSEA, CFI, AIC, BIC, and sample-size adjusted BIC) and attending to substantive concerns when adjudicating between models.

Given the widespread use of the EATQ-R and the potential for future research to implement this proposed scoring method, it is crucial to test the authors' claim that the findings of Snyder et al. (2015) are "robust and likely to generalize" (p. 1141). The present study seeks to examine the generalizability of the latent factor structure and concurrent validity findings of Snyder et al. (2015) in a large sample of Mexican-origin youth. Additionally, the present study aims to extend Snyder et al.'s findings by examining not only concurrent validity but also the predictive validity of the latent EATQ-R factors, using outcomes that allow us to evaluate both convergent and discriminant validity.

Replicability and Generalizability

Snyder and colleagues (2015) tested the replicability of their results by deriving the bifactor models using 50% of their sample and then examining the fit of these derived models in the remaining 50% of their sample (i.e., a randomly selected hold-out sample). This approach demonstrates, importantly, that the model fit of the latent structure of the EATQ-R replicates in a sample with the same characteristics as the derivation sample. However, it provides little insight into the generalizability of the findings, or whether the results depend on an originally unmeasured variable (e.g., ethnicity; Asendorpf et al., 2013). Generalizability is especially important for claims about construct validity as the degree to which a measure demonstrates construct validity varies by the specific use of the scale and "can often be context or population dependent" (Flake et al., 2017, p. 371). Because Snyder et al. (2015) found evidence for replicability, and replicability is necessary, but not sufficient, for generalizability, the next logical step is to examine the generalizability of the original findings in a sample that differs from the original sample in a measurable way (Finkel, Eastwick, & Reis, 2017).

In the present study, we examine the generalizability of both the derived latent factor model fit and associations with adolescent functioning in a sample of Mexican-origin youth. This study will help calibrate the theoretical conclusions drawn in Snyder et al. (2015) about the construct validity of the EATQ-R and facilitate the ongoing process of understanding how to best characterize adolescent temperament (Flake et al., 2017; Tackett & Durbin, 2017).

The Present Study

Our goal for the present study was to examine the generalizability of the findings from Snyder et al. (2015). We expected to replicate Snyder et al.'s (2015) findings in a sample of Mexican-origin youth, thereby demonstrating the generalizability of the findings to an ethnic

minority sample. Specifically, we expected to find: (1) adequate model fit (defined below) for the EC and NE bifactor models presented in Snyder et al. (2015); (2) adequate fit for a modified PE bifactor model[5]; (3) associations between the EC, NE, and PE latent factors (as specified by Snyder et al., 2015) and measures of adolescent functioning that are in the same direction, and of similar magnitude, as those found in Snyder et al. (2015); and (4) evidence that an alternative scoring method tested by Snyder et al. (2015) shows weaker and less specific associations with adolescent functioning measures than the latent factors.

In addition to testing these hypotheses, we also extended Snyder et al. (2015) in several exploratory ways. First, we examined the influence of correlated residuals on model fit for the retained bifactor models. Second, given the methodological limitations of the original study, we conducted new model comparisons to see whether the bifactor models were a better conceptual and empirical fit to the data than three competing models (single factor, correlated factors, hierarchical factor). Third, given research documenting gender differences in temperament (Else-Quest et al., 2006), we tested for measurement invariance across gender in the structure of adolescent temperament. Fourth, we examined whether the derived latent factors predict adolescent functioning assessed two years after the EATQ-R data were collected. In other words, we extended Snyder et al. (2015) by examining predictive, as well as concurrent, validity.

Method

Participants

The present study used data from the California Families Project, a longitudinal study of Mexican-origin youth and their parents ($N = 674$).[6] Children were drawn at random from rosters of students from the Sacramento and Woodland, CA school districts, in 2006–07. The focal child had to be in the 5th grade, of Mexican origin, and living with his or her biological mother, in order to participate in the study. Approximately 72.6% of the eligible families agreed to participate in the study, which was granted approval by the University of California, Davis Institutional Review Board (Protocol # 217484–21). The children (50% female) were assessed annually from 5th grade to two years post-high school. To most closely match the age of participants in Snyder et al. (2015), the present study used data from Wave 3, when the children were in 7th grade ($M_{age} = 12.81$, $SD = 0.49$). To extend the analyses from Snyder et al. (2015) to include an examination of predictive validity, the

[5]The present study does not have data from two PE subscales used in Snyder et al. (2015). Therefore, we fit a model where Surgency and Affiliation form two separate factors, rather than one general factor. This is consistent with findings from Snyder et al. (2015), which showed that Surgency and Affiliation share little common variance, and also follows the approach recommended by the first author of the original article (Dr. Hannah Snyder, personal communication, May 10, 2019).

[6]Nine published papers have used data from the California Families Project (CFP) to examine temperament measured via the EATQ-R (Atherton, Lawson, Ferrer, & Robins, 2020; Atherton, Lawson, & Robins, 2020; Atherton, Schofield, Sitka, Conger, & Robins, 2016; Atherton, Tackett, Ferrer, & Robins, 2017; Atherton, Zheng, Bleidorn, Robins, 2020; Clark, Donnellan, Conger, & Robins, 2015; Clark, Donnellan, & Robins, 2018; Robins, Donnellan, Widaman, & Conger, 2010; Taylor, Widaman, Robins, 2018). Some of these studies have computed correlations between EATQ-R domains and the adolescent functioning measures from the present study. Specifically, Atherton et al. (2017) examined how Effortful Control, Negative Emotionality, and Positive Emotionality were associated with relational aggression and used data overlapping with the present study at two out of four waves. Additionally, Atherton et al. (2020) examined how Effortful Control was associated with ADHD symptoms with data overlapping with the present study at two out of four waves. Importantly, these papers differ significantly from the present study in terms of research questions, scoring of the EATQ-R, and type of analyses. Moreover, no previous CFP publications have examined the association of any EATQ-R temperament domain with any of the other measures of adolescent functioning included in Snyder et al. (2015). For a full list of California Families Project publications, see: https://osf.io/ky7cw/.

present study also used data from Wave 5, when the children were in 9th grade (M_{age} = 14.75, SD = 0.49). The retention rate relative to the first assessment (Wave 1, 5th grade) was 86% at Wave 3 (N = 586) and 90% at Wave 5 (N = 605).

Participants were interviewed in their homes in Spanish or English, depending on their preference. Interviewers were all bilingual and most were of Mexican heritage. Sixty-three percent of mothers and 65% of fathers had less than a high school education (median = 9th grade for both mothers and fathers); median total household income was between $30,000 and $35,000 (overall range of income = < $5,000 to > $95,000). With regard to generational status, 83.6% of mothers and 89.4% of fathers were 1st generation, and 16.4% of mothers and 10.6% of fathers were either 2nd or 3rd generation. One hundred and twenty-four of the families were single-parent households (mothers only), and 549 of the families were two-parent households.

Compared to the participants in Snyder et al. (2015), participants in the present study differ in several ways (see Table 1). Notably, the majority of participants in the combined sample in Snyder et al. (2015) were White and lived in various geographic locations (i.e., Washington, Colorado, and New Jersey, United States; The Netherlands; Belgium) whereas all participants in the present study were Mexican-origin and living in northern California.

Measures

The measures described in the present study are identical to those used in Snyder et al. (2015) unless otherwise indicated.

Temperament—Adolescent temperament was measured via adolescent self-reports using the *Early Adolescent Temperament Questionnaire – Revised* (EATQ-R; Ellis & Rothbart, 2001).[7] The EATQ-R was designed to measure three domains of temperament – EC (M = 3.02, SD = 0.40, α = .74, ω = .72), NE (M = 1.96, SD = 0.38, α = .84, ω = .87), and PE (M = 2.70, SD = 0.40, α = .55, ω = .66). Descriptive statistics for all subscales are shown in Table S1.

EC Subscales: The EC scale (16 items) has three facets: Activation Control (5 items), Attention (6 items), and Inhibitory Control (5 items). Activation Control assesses the ability to perform an action or pursue goals when there are competing desires. Attention assesses the ability to focus and shift attention when needed. Inhibitory Control assesses the ability to plan and suppress inappropriate impulses.

NE Subscales: The NE scale (17 items) has three facets: Fear (6 items), Frustration (7 items), and Shyness (4 items). Fear assesses unpleasant affect derived from anticipating distress. Frustration assesses negative affect related to ongoing tasks being interrupted. Shyness assesses behavioral avoidance of novelty and social challenges. Aggression (6

[7]Participants completed the EATQ-R in either English (90%) or Spanish (10%). Researchers have translated a number of Rothbart's temperament measures into Spanish (e.g., González Salinas et al., 1999; González Salinas et al., 2000), including the EATQ-R (see https://research.bowdoin.edu/rothbart-temperament-questionnaires/instrument-descriptions/the-early-adolescent-temperament-questionnaire/). There is work examining the validity of the EATQ-R for the Chilean Spanish version (Hoffmann et al., 2017) and the Spanish version in a sample of Catalan-speaking Spanish adolescents (Viñas et al., 2015).

items) assesses hostile reactivity to negative stimuli including person- and object-directed violence. Depressed Mood (6 items) assesses unpleasant affect, lowered mood, and lack of enjoyment in activities. The Aggression and Depressed Mood scales also fall into the NE scale, as shown by Snyder et al. (2015).

PE Subscales: The PE scale (11 items) has two facets: Surgency (6 items) and Affiliation (5 items). Surgency assesses pleasure derived from activities involving high intensity or novelty. Affiliation assesses the desire for warmth and closeness with others. Notably, the PE measure in Snyder et al. (2015) also included four Perceptual Sensitivity items (assessing awareness of low-intensity stimulation in the environment) and five Pleasure Sensitivity items (assessing pleasure related to activities or stimuli involving low intensity), which were not assessed in the present study. Given the absence of these Perceptual and Pleasure Sensitivity items in the present study, we limit our findings about the EATQ-R PE domain to Surgency and Affiliation.

Adolescent functioning—Consistent with Snyder et al. (2015), adolescents reported on their depression, anxiety, and ADHD symptoms, as well as aggression towards and victimization by peers, whereas mothers completed reports of their child's school behavior and performance. All adolescent functioning measures were assessed at both age 13 (for concurrent validity analyses) and age 15 (for predictive validity analyses).

Depression, Anxiety, and ADHD: To assess depression, anxiety, and ADHD symptoms, the present study used the NIMH Diagnostic Interview Schedule for Children-IV (*DISC-IV*). The *DISC-IV* is a comprehensive, psychiatric interview that assesses mental health problems for children and adolescents using DSM-IV criteria; it is the most widely-used mental health interview that has been tested in both clinical and community populations and validated in both English and Spanish (Costello, Edelbrock, & Costello, 1985; Schwab-Stone et al., 1996; translated into Spanish by Bravo, Woodbury-Farina, Canino, & Rubio-Stipec, 1993). For the present study, we used the Depression (22 items), Anxiety (14 items), and Attention-Deficit Hyperactivity Disorder (ADHD; 24 items) modules of the NIMH *DISC-IV*. Responses were recorded dichotomously (0 = *no*, 1 = *yes*) as the symptom being present or not in the past year. The Depression module included questions about feeling sad and irritable such as, "[Was there] a time in the past year when you were very upset or depressed?" and physical symptoms of depression such as, "[Did you] sleep more during the day than usual in the last year?". The Anxiety module included questions about general worry and concern such as, "[Are you the] type of person who is tense and finds it hard to relax?" and physical symptoms of anxiety such as, "[Did you] often have stomachaches in the last year?". The ADHD module included questions about attention-related behaviors such as, "[Did you have] trouble keeping your mind on task for more than a short period of time?" and hyperactivity problems such as, "[Did you] often climb on things/run around when you weren't supposed to?". For Depression, Anxiety, and ADHD, we computed a symptom count variable by summing the responses for each symptom (present vs. absent) to create separate composite scores of Depression, Anxiety, and ADHD. In addition, we computed symptom counts for the inattention (11 items) and hyperactivity (12 items) facets of ADHD.

Although the *DISC-IV* is similar to the measures Snyder et al. (2015) used to assess psychopathology, it is important to note that the DISC-IV generates symptom counts whereas the Snyder et al. measure used Likert-type continuous rating scales of Depression (27-item self-report Children's Depression Inventory; Kovacs, 1985), Anxiety (39-item Manifest Anxiety Scale for Children; March, Parker, Sullivan, Stallings, & Conners, 1997), and ADHD (18-item MTA SNAP-IV; Swanson et al., 2001). Consequently, we expected to have more zero-inflated Depression, Anxiety, and ADHD scores than Snyder et al. (2015), as discussed in personal communication with the original first-author (personal communication, May 10, 2019), which would likely result in smaller effect sizes.

Antisocial behavior toward peers and victimization by peers: Adolescents completed a 12-item Relational Aggression scale (for details, see Aizpitarte et al., 2018; Atherton et al., 2017) that includes all seven items from the shortened Revised Peer Experiences Questionnaire (Prinstein, Boergers, & Vensberg, 2001) used by Snyder et al. (2015) to measure relational aggression. Participants completed two versions of the scale, one where they were asked about being a *victim* of relational aggression and the other where they were asked about being a *perpetrator* of relational aggression. Sample items include, "In the past three months, a kid your age told mean stories or lies about you." ["In the past three months, you told mean stories or lies about a kid your age."] and "In the past 3 months, a kid your age left you out of what he or she was doing." ["In the past 3 months, you left a kid your age out of what you were doing on purpose."]. Responses were made on a 4-point scale ranging from 1 (*almost never or never*) to 4 (*almost always or always*) and scores were calculated by averaging item responses for each version of the scale (i.e., mean victim score and mean perpetrator score for each participant).

School grades: School performance was measured using a one-item parent report assessing school grades, "On average, what are [ADOLESCENT'S] grades?" and the response scale ranging from 1(*mostly F's*) to 5 (*mostly A's*). This is the same item from Snyder et al. (2015), in which "Parents reported on their child's typical letter grades, from 1 = *mostly A's* to 5 = *mostly F's*" (p. 1136).

School behavior: School behavior was also assessed using a one-item parent report assessing adolescent misbehavior at school, "In the past 12 months, how frequently has [ADOLESCENT] been in trouble at school for things like arguing and fighting, being very disruptive in class, or other things like these?" and the response scale ranged from 1 (*never*) to 4 (*often*). This item is similar (but not exactly the same) as the item from Snyder et al. (2015), in which "Parents … reported the number of times their child had been sent to the office for misbehavior during the year from 1 = *none* to 6 = *more than five times*" (p. 1136).

Gender—At age 10, youth reported their gender (1 = girl; 2 = boy).

Data Analyses

All analyses were conducted in R version 3.6.0 (R Core Team, 2019) and RStudio Version 1.2.1335 using the *lavaan* (Rosseel, 2012) and *psych* packages (Revelle, 2018).[8] We used maximum likelihood (ML) estimation for all models, unless otherwise noted.

Replication of Snyder et al. EATQ-R factor structure analyses—Our first set of analyses focused on evaluating the fit of the latent structure of adolescent temperament reported in Snyder et al. (2015). Unless otherwise noted, the statistical methods detailed in this section are the same as those described in the Methods section of Snyder et al. (2015). Specifically, we tested the fit of the EC, NE, PE, and full bifactor models reported in Snyder et al. (2015) (Table S2 in the present study). To avoid overfitting to our data and to provide an accurate depiction of the generalizability of the findings from the original paper, no modifications were made from the original EC and NE bifactor models. As detailed above, we tested a modified PE bifactor model given that the present study excludes Perceptual Sensitivity and Pleasure Sensitivity scales. To fit the bifactor models, we used the "cfa" function in *lavaan*. Factor variance was set to 1 so that all factor loadings were estimated. Additionally, item loadings were standardized with respect to latent variable variance (i.e., std.lv = TRUE in *lavaan*). The same correlated residuals that were included in Snyder et al. (2015) were specified in our models (see Figures S1 – S3 for path diagrams).

Consistent with the thresholds outlined in the original article, absolute model fit was assessed as good if RMSEA < .05 and CFI > .95 and adequate if RMSEA < .08 and CFI > .90. Given the lack of consensus about rules of thumb for model fit indices (Marsh, Hau, & Wen, 2004) and complexities with interpreting model fit indices in factor analyses of personality measures (Hopwood & Donnellan, 2010), we report exact values for all fit indices. We also report (see Supplemental Materials, Tables S2–S3) several additional statistical indices for the bifactor models recommended by Rodriguez, Reise, and Haviland (2015) (see also Dueber, 2017), including omega hierarchical/omega hierarchical subscale, explained common variance (ECV), percent of uncontaminated correlations (PUC), item common variance attributed to a general dimension (I-ECV), a measure of construct reliability (i.e., H), and a measure of factor score determinacy (i.e., FD). In addition, we examined the congruence between the item-level factor loadings reported by Snyder et al. (2015) with those observed in the present study; that is, when rank-ordered by magnitude of the factor loading for each temperament domain (i.e., EC, NE, PE), how similar is the order of items between Snyder et al. and the present study. To do this, we calculated Pearson correlations between factor loadings from Snyder et al. and the present study.

Correlations among latent factors: After fitting the bifactor models for each domain, we estimated a full model with all three domains (EC, NE, and PE) together. This model allowed us to examine associations among all latent factors, both within (e.g., NE Fear-specific factor with NE Frustration-specific factor) and across temperament domains (e.g., NE Fear-specific factor with EC Activation Control-specific factor).

[8]Our project attempted to replicate and extend the analyses from Snyder et al. (2015). Given our primary goal of replication, we did not preregister our initial analyses, and instead used the exact same methods from Snyder et al. (2015) to replicate their work. During the publication process, the reviewers and editor recommended that we conduct additional analyses. These analyses are preregistered on the project OSF page, which also includes R scripts to run all analyses and materials: https://osf.io/5rhjb/. We made one deviation from our supplemental preregistration. Specifically, when we conducted new model comparisons, we considered the interpretability of the factor loadings (e.g., whether all items loaded in the expected direction), in addition to model fit indices, to adjudicate between models. Additionally, the CFP participants have not given informed consent to have their personal data publicly shared, and we do not have IRB approval to post data publicly. Therefore, we are legally and ethically not allowed to publicly post CFP data. Researchers interested in replicating findings can contact the authors to gain access to individual-level data. Further, we are unable to share full materials for the NIMH DISC-IV depression, anxiety, and ADHD scales as they are copyrighted psychiatric assessments.

Model comparisons for different EATQ-R factor structures—Next, we attempted to account for limitations in the methods used by Snyder and colleagues to determine the best-fitting model of the factor structure of the EATQ-R in the present sample. Given that bifactor models can be erroneously identified as the best-fitting model when a more parsimonious model fits similarly well (e.g., Morgan et al., 2015), we compared a number of different CFA models. In particular, for each temperament domain (i.e., EC, NE, PE), we examined (1) a single-factor model where all of the items load onto a single domain factor, (2) a correlated-factors model where items load onto their respective subscale and then these latent subscale scores are allowed to correlate, (3) a bifactor model without correlated residuals and without excluding any EATQ-R items, and (4) a hierarchical model with one higher-order factor (e.g., EC) and the relevant subscales (e.g., activation control, attention control, and inhibitory control) as specific factors subsumed within the higher-order factor. Unlike Snyder et al. (2015), we did not remove items that load below .30, as doing so could reduce construct validity because items with low loadings may tap into relatively unique content that is nonetheless theoretically important to the construct. Two of these models (i.e., single-factor and correlated-factors models) were tested in Snyder et al. (2015) and deemed to exhibit worse fit than their retained bifactor model. However, Snyder and colleagues used change in chi-square as a primary method to compare model fit, which is problematic in large samples because even trivial differences in fit can be statistically significant (Cheung & Rensvold, 2002). Thus, to better compare the models (i.e., single-factor, correlated-factors, bifactor, hierarchical) in terms of model fit, we used the following indices – RMSEA, CFI, AIC, BIC, and sample-size adjusted BIC. We consider the best-fitting model to be the one that had the best fit in the majority of these five indicators. We also considered substantive concerns when adjudicating between models – in particular, having interpretable factor loadings (e.g., items loading in the expected direction) (Morgan et al., 2015). We then conducted all subsequent analyses using the best-fitting model from our model comparisons and the retained bifactor model from Snyder et al.[9]

Measurement invariance—As an extension of Snyder et al., we examined measurement invariance across gender. To do this, we compared a series of multiple group models. In particular, we examined four models: (a) freely estimating the factor loadings across for boys and girls (i.e., configural invariance); (b) constraining the respective factor loadings to be equal for boys and girls (i.e., weak invariance); (c) constraining the factor loadings and intercepts for boys and girls (i.e., strong invariance); and (d) constraining the factor loadings, intercepts, and residual variances for boys and girls (i.e., strict invariance). If the more constrained models did not fit worse than the lesser constrained models, then we concluded that the structure of the adolescent temperament is similar for boys and girls.

Correlations with adolescent functioning—To address questions about concurrent validity, we examined the concurrent correlations between the latent temperament factors and each measure of adolescent functioning. As an extension of Snyder et al. (2015), we also examined the predictive validity of the latent temperament factors by examining

[9]For NE, we ran analyses both with and without the Depressed Mood and Aggression subscales. To be consistent with Snyder et al. (2015, p. 1138, Footnote 6), we report results from the analyses including Depressed Mood and Aggression in the main text, and we report results excluding Depressed Mood and Aggression in the supplemental materials (see Tables S12 – S13).

the correlations between the latent temperament factors at age 13 and each measure of adolescent functioning at age 15. For both concurrent and predictive validity analyses, we used the "rcorr" function in the *Hmisc* package (Harrell, 2019). For zero-inflated count data (i.e., psychiatric symptoms), we used Spearman's *rho* correlations and for all other adolescent functioning variables, we computed Pearson's *r* correlations.

Alternative scoring method—In addition to examining concurrent and longitudinal correlations between the derived latent factor structure and adolescent functioning outcomes, we also examined the concurrent and longitudinal associations between adolescent functioning outcomes and the alternative scoring method for the EATQ-R presented in Snyder et al. (2015). This alternative scoring method (also called the "traditional method" by Snyder et al.) differs from the other scoring methods examined in this study in two ways. First, the content included in each domain differs, such that EC is a composite of Attention, Activation Control, and Inhibition; NE is a composite of Fear, Aggression, Frustration, and Shyness (without Depressed Mood); and PE is a composite of Surgency and Affiliation (given that the present study excludes Pleasure Sensitivity and Perceptual Sensitivity). Second, this alternative scoring method uses observed (or manifest) composites for each domain rather than latent variables. Given that latent variables are free of non-systematic measurement error, correlations between observed variables should be smaller (Bollen, 2002; Borsboom, 2008). The results of all analyses examining this alternative scoring method are provided in the Supplemental Materials (see Table S8).

Power analyses—Power analyses for the correlations between derived latent factors and measures of adolescent functioning were performed using the *pwr* package (Champely, 2018). Using our sample size of 586 and setting alpha = .05 with a two-sided alternative hypothesis, we had 90% power to detect an effect size of $r = .13$ and 95% power to detect an effect size of $r = .15$. This was sufficient power to detect all significant correlations in Snyder et al. (2015) (see italicized values in Table 7).

Results

Replication of Snyder et al. EATQ-R Factor Structure Analyses

We fit the bifactor models from Snyder et al. (2015) to each of the three temperament domains. Table 2 shows bifactor model fit statistics from the present study. Factor loadings and I-ECV for each item are shown in Table S2. Additional statistical indices (i.e., omega, omega hierarchical, ECV, PUC, H, and FD) are reported in Table S3.

Effortful Control—For EC, Snyder et al. found that the best-fitting latent factor model included a common EC factor and an Activation Control-specific factor, with no Inhibitory Control-specific or Attention-specific factors. Additionally, we removed item 41 ("You are good at keeping track of several different things that are happening around you") because it was excluded from Snyder et al.'s EC model for its weak loading (i.e., below .30) on the Attention subscale. When we fit this EC model to our data, we found that the model had poor fit by both CFI and RMSEA. However, the χ^2-value is almost identical to the one from the hold-out replication sample in Snyder et al. When we compared the factor loadings

between the original study and the present study, we found a correlation of $r = .31$ for the Common EC factor and $r = -.22$ for the Activation-specific factor.

Negative Emotionality—For NE, Snyder et al. (2015) found that the best-fitting latent factor model included a common NE factor and specific factors for all subscales (i.e., Fear, Frustration, Shyness, Aggression, and Depressed Mood). In the original article, item 37 ("You get sad when a lot of things are going wrong") had a weak negative loading on Depressed Mood and was eliminated from the Depressed Mood-Specific factor (but not the common NE factor), so we also loaded this item only on the common NE factor. When we fit this NE model to our data, we found that the model had poor fit via CFI and adequate fit via RMSEA (Table 2). Notably, the CFI and RMSEA fit indices are quite similar to the original study for the NE domain. When we compared the factor loadings between the original study and the present study, we found a correlation of $r = .58$ for the Common NE factor, $r = -.12$ for the Depressed Mood-specific factor, $r = .42$ for the Fear-specific factor, $r = .48$ for the Shyness-specific factor, $r = .73$ for the Frustration-specific factor, and $r = .87$ for the Aggression-specific factor.

Positive Emotionality—For PE, Snyder et al. (2015) found that the best-fitting latent factor model included a common PE factor, specific factors for Affiliation, Perceptual Sensitivity, and Pleasure Sensitivity, and Surgency as a separate factor (not in the common PE factor). Given that the present study only included data on Affiliation and Surgency, we only included these two factors when fitting the PE model. In particular, we fit a model where Affiliation and Surgency formed two specific factors that were allowed to correlate, which is consistent with findings from Snyder et al. (2015) and with recommendations from the first-author (personal communication, May 10, 2019). Additionally, we removed two Surgency items (item 3 "You think it would be exciting to move to a new city" and reverse-scored item 19 "You wouldn't like living in a really big city, even if it was safe") because they were excluded from Snyder et al.'s PE model for their weak loadings (i.e., below .30) on the Surgency subscale. When we fit this modified PE model to our data, we found that the model had poor fit via CFI and adequate fit via RMSEA (Table 2). When we compared the factor loadings between the original study and the present study, we found a correlation of $r = .92$ for the Surgency factor and of $r = .72$ for the Affiliation factor.[10]

Full combined model—Snyder et al. (2015) estimated a full model that included the EC, NE, and PE domains modeled together. Our full model was modified because it included the modified PE factors. When we fit this modified full model to our data, we found that the model had poor fit via CFI and adequate fit via RMSEA (Table 2), similar to Snyder et al.'s findings for their full model.

[10]We also ran additional exploratory analyses examining whether model fit was improved when we used an estimator that is appropriate for *ordinal* data, the diagonally weighted least squares with mean and variance adjustments (WLSMV) estimator. In contrast to WLSMV, ML estimation relies on *interval*-level indicators, but the Likert scales used to rate the EATQ-R items arguably fail to attain interval-level measurement, which could contribute to the suboptimal model fit we found for many models using ML. Consistent with this reasoning, we found substantially better model fit for EC and NE (but not PE) when we used WLSMV estimation instead of ML estimation (Table S4).

This full model allowed us to examine associations between all of the latent EC, NE, and PE factors. The original article found that EC was negatively correlated with most NE factors and that Surgency was negatively correlated with the Activation Control-Specific factor, the Fear-Specific, and the Shyness-Specific factor. Based on significance tests, these findings from Snyder et al. (2015) did not generalize to the present sample, as we found far fewer significant correlations (Table 3). Indeed, only 2 out of 14 (14%) significant correlations replicated between the original and present studies; in both studies there were significant negative associations of the common EC factor with the specific Aggression factor and the specific Depressed Mood factor. However, if we consider only the direction of the effect, we did find that 18 out of 22 (82%) of the correlations were in the same direction across the two studies.

Further, we found a number of significant associations between the Affiliation-specific factor and NE and EC factors. With the caveat that the Affiliation-specific factor was specific to our study and cannot be directly compared to results from the original study, the pattern of correlations does parallel those found by Snyder and colleagues (2015) for the common PE factor in terms of direction and significance.

Exploratory Bifactor Model Analyses

Given the generally suboptimal fit of the EC, NE, PE, and full bifactor models, we ran exploratory analyses to assess whether the model deficiencies were due, at least in part, to correlated residuals that were not being modeled. More specifically, we examined (1) if the correlated residuals from Snyder et al. (2015) actually increased model fit in the present study, and (2) whether adding new correlated residuals might lead to even better model fit in the present study. We believe these exploratory analyses are needed to thoroughly evaluate the generalizability of the findings from Snyder et al. (2015) because they added item residual covariances “until good model fit was achieved” (p. 1136). In other words, item covariances were included based on modification indices that reflect purely empirical (rather than theoretical) associations. Consequently, this practice is likely to capitalize on sample-specific associations and may decrease generalizability of the model when fit to other datasets.

First, to examine whether the correlated residuals from Snyder et al. (2015) actually increased model fit in the present study, we re-ran bifactor models with these residual covariances removed and compared fit indices to those of the original model. As another comparison, we also conducted analyses where we added correlated residuals between *random* pairs of items. (We ran these analyses three times, each with unique random pairs of items, and averaged the results across these three trials.) We found that the correlated residuals from Snyder et al. (2015) led to better model fit than not having any correlated residuals or having random correlated residuals, as the CFIs were higher and the RMSEAs were lower when fitting the original model versus the model without any correlated residuals and random correlated residuals (Table 4). Together, these results suggest that the data-dependent model modifications in Snyder et al. (2015) provide some generalizable benefit across samples.

Second, we examined whether there were different residual covariances that would result in better model fit for our data than the ones included in Snyder et al. (2015). To do this, we used the models with no correlated residuals and examined modification indices for residual covariances that would result in the greatest increase in model fit (i.e., reduce discrepancies between the observed and model-implied matrices). For consistency with Snyder et al. (2015), we included the same number of residual covariances that were included in their retained models (i.e., 4 residual covariances to EC, 6 residual covariances to NE, and 1 residual covariance with PE). Results indicate that the model with modified correlated residuals resulted in better fit for the EC and NE subscales and equal fit for the PE subscale.[11] These findings are not surprising, given that these model modifications were made via purely empirical (rather than theoretical) decisions. Further, the results highlight how data-dependent model modifications may generalize across multiple studies but they also might provide less benefit outside of a particular sample.

New Model Comparisons

Next, we conducted additional model comparisons examining single-factor, correlated factors, bifactor, and hierarchical factor models. Results from the new model comparisons for each of the temperament domains (EC, NE, and PE) are shown in Table 5. For EC, we found that the bifactor model fit the best given the statistical fit indices.[12] However, in some cases, the factor loadings in this model did not correspond to the presumed conceptual meaning of the factor. In particular, four items (two from the Attention-specific factor and two from the Inhibitory Control-specific factor) did not load in the expected direction. For example, the positively keyed item, “You can stick with your plans and goals”, had a *negative* loading (−.15) on the Inhibitory Control-specific factor, making it difficult to interpret this factor as reflecting high levels of inhibitory control. For NE, the correlated factors model fit as well as (or better than) the bifactor model according to all of the fit indices, and better than both the single factor and hierarchical models. For PE, the bifactor model fit best based on the fit indices. However, this model was not positive definite because of a negative variance. The next best model was the correlated factors model, which fit slightly better than the hierarchical model.[13]

Overall, these model comparisons did not clearly identify a best-fitting model. For NE, the correlated factors model fit the best and is the most parsimonious. However, for EC and PE, the bifactor model fit best according to the statistical fit indices, but had some confusing factor loadings for EC and was not positive definite for PE. Given that the correlated factors model was the best or second best model (considering both empirical and conceptual issues) for all three domains, and given that we are already examining the bifactor models

[11]Items that had correlated residual variances for the original and modified models can be found in Table S2. For the EC scale, none of the pairs of item residual covariances overlapped between Snyder et al. (2015) and the present study. For the NE scale, one pair of item correlated residuals overlapped (i.e., residual variances between item 60 and 64). For the PE scale, the same single pair of items had correlated residuals in both Snyder et al. (2015) and the present study, resulting in identical model fit for both the original model and the modified correlated residuals model.

[12]For EC, the hierarchical model did not converge. Additionally, in the correlated factors model, the correlation between the Attention and Inhibitory Control subscales had to be constrained to 1.

[13]We also ran these model comparisons using WLSMV estimation and found a similar pattern of findings, but with substantially better model fit for all models (Table S5). In other words, our conclusions about which model fit best were the same regardless of whether we used ML or WLSMV estimation.

from Snyder et al., we decided to retain the correlated factors models from our model comparisons for EC, NE, and PE (see Figures S4–S6 for path diagrams).

Measurement Invariance across Gender

Snyder et al. bifactor models—Results of the measurement invariance analyses for the retained bifactor models from Snyder et al. are shown in Tables S6. We found evidence for strict invariance for EC, strict invariance for NE, and weak invariance for PE.

Correlated factors models—Results of the measurement invariance analyses for the correlated factors models are shown in Tables S7. We found no form of invariance for EC (the configural invariance model did not converge), weak invariance for NE, and weak invariance for PE.

Concurrent and Prospective Associations with Measures of Adolescent Functioning

Snyder et al. bifactor models—Next, we used the bifactor models from Snyder et al. (2015) to examine the concurrent associations between the derived temperament scores and measures of adolescent functioning. Descriptive statistics (i.e., means, standard deviations, ranges) for all measures of adolescent functioning are shown in Table 6. 95% confidence intervals for correlations are presented in Table S14.

Correlations of Effortful Control (EC) with adolescent functioning: Snyder et al. (2015) found that higher common EC scores were associated with fewer concurrent symptoms of depression, anxiety, and ADHD; less antisocial behavior toward peers and victimization by peers; better grades; and fewer disciplinary problems. On the whole, these findings generalized to the present sample (Table 7), with two exceptions; using the Bonferroni adjusted alpha level of .0003 from Snyder et al. (2015), we did not find a significant association between common EC scores and lower levels of victimization by peers ($r = -.11$, $p = .006$) or the common EC scores and fewer symptoms of anxiety ($r = -.13$, $p = .002$). Therefore, the vast majority of the concurrent associations found by Snyder and colleagues replicated in the present sample.

As an extension of Snyder et al. (2015), we examined associations between EC scores at age 13 and measures of adolescent functioning two years later (Table 7). We found similar, but weaker, prospective associations related to those associations we found concurrently. In particular, we found that that higher common EC scores at age 13 were associated with fewer ADHD symptoms ($r = -.20$, $p < .001$), less antisocial behavior toward peers ($r = -.18$, $p < .001$), better grades ($r = .23$, $p < .001$), and fewer disciplinary problems at age 15 ($r = -.22$, $p < .001$). We did not find any significant prospective associations between the Activation Control-specific factor and any of the adolescent functioning measures.

Correlations of Negative Emotionality (NE) with adolescent functioning: Snyder et al. (2015) found that higher common NE scores were associated with more symptoms of depression and anxiety, and more antisocial behavior toward peers and victimization by peers. These findings all generalize to the present sample (see Table 7). Additionally, we also found a positive significant association with common NE and ADHD ($r = .28$, $p <$

.001). Snyder et al. (2015) found that higher Aggression-specific scores were associated with more antisocial behavior toward peers and victimization by peers, lower grades, and more school discipline problems. These associations generalize to the present sample, with the exception that we did not find evidence for higher Aggression-specific scores and more victimization by peers ($r = .07$, $p = .076$). In addition to the replicated findings, we also found that youth with higher Aggression-specific scores also had more ADHD symptoms ($r = .30$, $p < .001$). Further, Snyder et al. (2015) found that higher Depressed Mood-specific scores were associated with more depression and physical anxiety symptoms. We replicated the positive association between Depressed Mood-specific scores and depression symptoms ($r = .27$, $p < .001$) and also found that higher Depressed Mood-specific scores were associated with more ADHD symptoms ($r = .19$, $p < .001$) and more victimization by peers ($r = .20$, $p < .001$). Snyder et al. (2015) found that higher Fear-specific scores were associated with more anxiety symptoms and lower grades. We replicated the significant positive association between Fear-specific scores and anxiety symptoms ($r = .18$, $p < .001$). Unlike in the original article, we found significant associations between Frustration-specific scores with depression and ADHD symptoms. Finally, as in Snyder et al. (2015), we also found no association between the Shyness-specific scores and any measure of adolescent functioning.

The prospective associations were similar to (but weaker than) the concurrent associations for the common NE and Aggression-specific scores, with two exceptions (see Table 7). Using the Bonferroni adjusted alpha level of .0003 from Snyder et al. (2015), there was no association between common NE scores at age 13 and antisocial behavior towards peers at age 15 ($r = .15$, $p = .001$), and no association between Aggression-specific scores at age 13 and depression symptoms at age 15 ($r = .04$, $p = .294$). Furthermore, there were no prospective associations between Depressed-mood specific, Fear-specific, Frustration-specific, Shyness-specific scores and any of the measures of adolescent functioning except for a positive correlation between Depressed-mood specific scores at age 13 and depression symptoms at age 15 ($r = .20$, $p < .001$).

Correlations of Positive Emotionality (PE) with adolescent functioning: Snyder et al. (2015) found that Affiliation-specific scores were associated with higher Depression and Anxiety symptoms, more antisocial behavior toward peers, and more victimization by peers. These findings did not generalize to the present sample, as we found no significant concurrent associations between Affiliation and measures of adolescent functioning (Table 7). Similar to Snyder et al. (2015), we also found no significant associations between Surgency scores and measures of adolescent functioning.

Finally, Surgency and Affiliation did not have any significant prospective associations with any measure of adolescent functioning, except for a positive correlation between Surgency at age 13 and ADHD symptoms two years later ($r = .18$, $p < .001$) and Affiliation at age 13 and school grades two years later ($r = .19$, $p < .001$).

Comparison with an alternative method of scoring the EATQ-R temperament factors: In addition to assessing the three temperament domains using Snyder et al.'s latent factors, we also used an alternative scoring of the temperament domains and examined

their concurrent and prospective associations with adolescent functioning. In particular, when Snyder et al. (2015) examined differences between their derived latent factors and the traditional scoring method, they found that the correlation patterns were much less specific and that contamination by common variance masked some specific effects. Results of these analyses are presented in Table S8. We found strong evidence of generalizability, with 41 out of 50 (82%) effects reported by Snyder et al. showing significant associations in the present study. In addition, we found 25 significant associations between manifest EATQ-R scores and concurrent measures of adolescent functioning that were not significant in Snyder et al. but were all in the same direction (Table S8).

The prospective associations showed a similar, but weaker, pattern (Table S8). Of the 41 concurrent associations from Snyder et al. that replicated in the present study, 30 associations (73%) also showed significant prospective associations in the present study.

Correlated Factors Models—Finally, we examined concurrent and prospective associations between factors scores from the correlated factors models with measures of adolescent functioning (Table 8; 95% confidence intervals are presented in Table S15). To help us interpret these results, and how they converge and diverge with results using the bifactor models, we present in Tables S9–S11 the correlations between the factor scores derived from the bifactor models with the factor scores derived from the correlated factors models.

Correlations of Effortful Control (EC) with adolescent functioning: The three EC factors (Activation Control, Inhibitory Control, and Attention) from the correlated factors model had associations with adolescent functioning that were very similar to those observed for the common EC factor from the Snyder et al. bifactor models. In particular, all three subscales were associated with symptoms of Depression and ADHD, interpersonal aggression, and school grades and discipline. Notably, using the correlated factors model, the Activation Control subscale factor showed the expected associations with school grades ($r = .30$, $p < .001$) and discipline ($r = -.18$, $p < .001$), whereas the Activation-specific factor from the Snyder et al. retained bifactor model did not (in either their study or the present study).

The correlations between the factors derived from the bifactor and correlated factors models provide some insight into why this discrepancy exists (see Table S9). In particular, there are very high correlations ($rs = .96$ - $.98$) between factor scores from the Common EC factor from the bifactor model and the three EC subscales from the correlated factors model. However, there is a relatively low correlation ($r = .15$) between the Activation-specific factor from the bifactor model and the Activation Control factor from the correlated factors model. Together, these findings suggest that the EC items are tapping into a shared effortful control factor resulting in similar associations across the subscales.

Correlations of Negative Emotionality (NE) with adolescent functioning: The Aggression, Depression, and Frustration factors from the correlated factors model showed very similar associations as the common NE factor from the bifactor model, especially for symptoms of Depression, Anxiety, and ADHD and interpersonal aggression and victimization. The Fear factor was associated with symptoms of Depression and Anxiety

and interpersonal victimization, but only the association with Anxiety overlaps with the Fear-specific factor from Snyder et al.'s bifactor model. The Shyness factor was associated with Anxiety symptoms, whereas the Shyness-specific factor was not. This suggests that, in the bifactor model, the common NE factor is capturing much of the variance that is associated with adolescent functioning measures, leaving the specific factors with relatively less valid variance. Conversely, in the correlated factors model, each subscale factor is able to correlate more strongly with the theoretically related measures of adolescent functioning because the shared variance among the subfactors has not been removed.

With regard to the correlations between the factors from the bifactor and correlated factors models (see Table S10), we found moderate to strong correlations (*r*s = .53 - .95) between the Common NE factor scores and scores from each of the subscales from the correlated factors model, as well as between each specific factor and its corresponding factor from the correlated factors model (*r*s = .44 to .90).

Correlations of Positive Emotionality (PE) with adolescent functioning: As with Snyder et al.'s bifactor models, the Surgency and Affiliation factors from the correlated factors models were not significantly associated with any of the adolescent functioning measures with two exceptions: the Surgency factor was prospectively (but not concurrently) associated with total ADHD symptoms ($r = .19$, $p < .001$) and the Affiliation factor was prospectively (but not concurrently) associated with school grades ($r = .20$, $p < .001$). This pattern is consistent with the near-perfect correlations (*rs* = .98 – 1.00) between the factor scores from the bifactor and correlated factors models (see Table S11). Thus, regardless of the method of scoring the PE domain items, PE as assessed by the EATQ-R was not significantly associated with the eight measures of adolescent functioning examined in the present study.

Discussion

The present study aimed to replicate and extend Snyder et al.'s (2015) findings using data from a large sample of ethnic minority youth. Overall, we found weak evidence for the generalizability of the bifactor models reported in the original study for the EATQ-R temperament domains, with relatively poor model fit observed for the EC domain and adequate, but not good, fit observed for the NE and PE domains. Further, when we conducted new model comparisons, we found that correlated factors models produced more interpretable results in our data, although the EC and PE models did not fit well by traditional standards for fit indices. Together, these results suggest that the EATQ-R does not have a clear and replicable internal structure. In contrast, we replicated most, but not all, of the concurrent associations between temperament and adolescent functioning, and showed that these associations hold up longitudinally when predicting adolescent functioning two years later. These concurrent and prospective associations support the construct validity of the EATQ-R as a measure of adolescent temperament, despite its structural problems. Below we review and discuss these findings, and then turn to broader implications and directions for future research.

Do the Bifactor Models of the EATQ-R Temperament Domains Replicate?

Effortful control—Given the conceptual cohesiveness of the EC domain (i.e., all three subscales scales are theorized to be interrelated facets of a superordinate EC domain), we expected to find adequate fit for the bifactor EC model reported by Snyder et al. (2015). However, this model fit poorly in the present sample and we found weak correlations between the magnitude of the item loadings in Snyder et al. and the present study for both the Common EC factor ($r = .30$) and Activation-specific factor ($r = -.22$), suggesting that the EC bifactor model findings reported by Snyder et al. do not generalize well to our sample.

There are a number of potential explanations for this lack of generalizability. First, the EC factor structure may not generalize to our sample due to differences in age, ethnicity, SES, geographic location, and/or nationality between the samples (see Table 1). Second, and related to the first issue, the poor model fit may be due, at least in part, to acquiescence bias, or the tendency for participants to endorse items without regard to the actual content. In particular, acquiescence bias might have been a problem for the EC subscale where we found strong positive correlations between items keyed in the same direction, but much weaker associations between positively and negatively keyed items, suggesting the presence of a keying factor. Indeed, past research has found that low SES participants are more likely to exhibit acquiescence bias (Meisenberg & Williams, 2008) and participants in the present study come from lower SES backgrounds than the Snyder et al. participants. Therefore, this is consistent with the possibility that participants in the present sample may have been more likely to acquiesce than those in the original study. Third, the EC bifactor models may not robustly generalize because of underlying issues with bifactor models (Morgan et al., 2015; Murray & Johnson, 2013; Watts, Poore, & Waldman, 2019). The likelihood of this explanation is bolstered by the fact that modified residual covariances resulted in substantially better model fit of the EC model than the one that included the original residual covariances.

However, for EC, we also were unable to find a good-fitting factor structure when we evaluated the fit of single-factor, correlated factors, and hierarchical factor models, in addition to bifactor models. These findings dovetail with previous discussions about how Rothbart's temperament measures were developed through a theory-driven, top-down approach with little psychometric work to ensure a coherent internal structure (e.g., Kim, Brody, & Murry, 2003; Kotelnikova et al., 2016; Kotelnikova et al., 2017; Latham et al., 2020; Muris & Meesters, 2009). Further, these findings are consistent with a warning from Hopwood and Donnellan (2010) that, due to the inherent complexity of personality data, failure to find adequate fit for CFA models of personality data is more of the rule than the exception. In fact, they found that, when examining seven prominent, well-validated personality measures using CFA, all had CFIs (ranging from .65 to .79) and RMSEAs (ranging from .09 to .13) that were not acceptable by traditional standards. In the present study, the model fit indices for the EC correlated factors model fit into these ranges (i.e., CFI = .66, RMSEA = .088). Further, these issues may be due, at least in part, to the fact that we fit CFAs to item-level data, rather than using composite scores or parcels. As Hopwood and Donnellan (2010) noted, personality items often "tap additional if substantially minor sources of variation", which result in general model misfit if many correlated residuals

are not included (p. 334). To address this issue, we ran an exploratory analysis using item parcels for the EC correlated factors model and found that model fit improved substantially (CFI = .93, RMSEA = .063). Another potential impediment to acceptable model fit derives from our use of ML estimation, which assumes interval-level data—a standard that personality test items arguably fail to reach. When we conducted the CFAs using an estimator (WLSMV) designed for use with ordinal data, we found improved model fit (see Table S4), especially for the EC model (CFI = .85, RMSEA = .077, vs. .72 and .084 using ML estimation). Together, these analyses indicate that CFA models of the structure of the EATQ-R can attain improved model fit through the use of item parcels or WLSMV estimation, otherwise researchers may have to rely on numerous correlated residuals to achieve adequate fit.

Negative emotionality—Unlike the EC model, we found that the NE bifactor model fit as well in the present sample as it did in Snyder et al.'s holdout replication sample, supporting the generalizability of the NE factor structure. In particular, we found nearly good fit via RMSEA (but not adequate fit via CFI) for the NE bifactor model. In addition, we found that the model fit improved only slightly when modified correlated residuals were included, which further supports Snyder et al.'s claim that NE is a conceptually coherent construct. Further, we found moderate to high correlations between the item loadings in Snyder et al. and the present study for the Common NE factor (r = .58) and the specific factors (Fear r = .72; Shyness r = .48, Frustration r = .73, and Aggression r = .87), except for Depressed Mood (r = −.12). In new model comparisons, we found that the correlated factors model fit slightly better than the bifactor model, and we found high correlations between the respective factor scores from both of these models. Finally, we found substantially improved – and good – model fit for both NE models using WLSMV estimation. Together, these findings suggest that we are closer to a generalizable factor structure of NE (as measured by the EATQ-R) than we are for the EC domain.

Positive emotionality—In contrast to the EC and NE domains, the implications of the results for PE are less clear. Snyder et al. did not find any coherent common factor for PE, and our findings do little to clarify the structure of this domain (as measured by the EATQ-R), in part because the present study included only the Surgency and Affiliation scales and omitted the Perceptual Sensitivity and Pleasure Sensitivity scales. Nonetheless, our findings do corroborate Snyder et al.'s finding that Surgency and Affiliation (along with Pleasure Sensitivity and Perceptual Sensitivity) form separate factors (see also Latham et al., 2020). Additionally, we found high correlations between the item loadings in Snyder et al. and the present study for both Surgency (r = .92) and Affiliation (r = .72). Further, our omission of the Pleasure Sensitivity and Perceptual Sensitivity may not have mattered much with regard to the structure of this domain, given that these scales likely would have loaded onto the Affiliation factor. In the new model comparisons, we retained the PE correlated factors model, which is practically identical to the PE bifactor model because in both cases there was no common factor shared between the Surgency and Affiliation subscales. This similarity is highlighted when examining the near-perfect correlations between the respective factor scores from both the retained Snyder et al. bifactor model and correlated factors model. Finally, it is worth noting that the Surgency construct (also called High

Intensity Pleasure by Rothbart and colleagues) is the most closely aligned with Rothbart's theoretical conceptualization of Positive Emotionality, as well as other conceptualizations of PE (e.g., Watson, Clark, & Carey 1988), which tend to emphasize high arousal positive affect and reward sensitivity (Rothbart & Ahadi, 1994). In contrast, the content assessed by the Affiliation scale, and the two omitted scales measuring general sensitivity to stimuli, do not seem like core components of the PE domain.

Full model—We obtained poor model fit when we attempted to replicate Snyder et al.'s full model, which included the EC, NE, and PE domains modeled together. Despite finding better fit using WLSMV estimation, it remained poor fit by CFI and only adequate by RMSEA. However, it is worth noting that Snyder et al. also found poor fit for the overall model in their own data, so it is not surprising that this model did not generalize well to a sample that is different in so many ways from their sample. Nevertheless, this full model allowed us to examine associations between all of the latent EC, NE, and PE factors. Likely due at least in part to the poor model fit and low correlations between factor loadings from both studies, we only replicated 14% of the significant correlations replicated between the latent factors (82% of the correlations were in the same direction across the two studies). Together, these results further highlight that the EATQ-R lacks a coherent internal structure.

Measurement invariance—For the EC domain, we found evidence for strict measurement invariance across gender for the Snyder et al. bifactor model, but no measurement invariance for the correlated factors model. This suggests that, for the Snyder et al. bifactor model, the measurement parameters (i.e., factor loadings, intercepts, residual variances) are similar for both boys and girls (Van De Shoot et al., 2015; Putnick & Bornstein, 2016). For the NE domain, we found strict measurement invariance across gender for the Snyder et al. bifactor model and weak invariance for the correlated factors model. This suggests that, for Snyder et al. bifactor model, the measurement parameters (i.e., factor loadings, intercepts, residual variances) are similar for both boys and girls, whereas for the correlated factors model, we can only conclude that each item contributes to the latent constructs to a similar degree for both boys and girls. For the PE domain, we found weak measurement invariance across gender, which suggests that, for both models, each item contributes to the latent constructs to a similar degree for both boys and girls.

Do the Associations between Temperament and Adolescent Functioning Replicate?

We found strong evidence for the generalizability of the vast majority of the associations between the three temperament factors and measures of adolescent functioning.

Effortful control—For the EC domain, we replicated 5 out of the 7 (71%) significant concurrent associations with adolescent functioning outcomes found in Snyder et al.; specifically, youth higher in EC tended to show lower levels of depression, ADHD, and relational aggression, and better school grades and school behavior. For the five replicated effects, the median effect size was .36 in Snyder et al. and .31 in the present study. Further, of the 5 significant concurrent associations that we replicated, all 5 (100%) replicated prospectively, suggesting that these EC domains are related to important adolescent functioning outcomes not only at the same age but also two years later.

Moreover, we found 6 additional significant concurrent associations that were not found in the original study. Specifically, youth higher in common EC had higher levels of both the Inattention and Hyperactivity facets of ADHD and youth higher in the Activation-specific factor of EC had fewer symptoms of depression and ADHD (mirroring associations of the common EC factor and measures of adolescent functioning).

Negative emotionality—We replicated 8 out of the 10 (80%) significant concurrent associations between NE and adolescent functioning found in Snyder et al. For the 8 replicated effects, the median effect size was .41 in Snyder et al. and .26 in the present study. Seven of the eight (88%) significant concurrent correlations replicated prospectively, suggesting that these NE domains maintain their implications for adolescent functioning over time.

Moreover, we found 14 additional significant concurrent associations that were not found in the original study. Specifically, youth high in common NE and specific Aggression, Depression, and Frustration scores reported more ADHD symptoms. Additionally, youth with higher Depression specific scores were more likely to report being victims of relational aggression. Finally, youth with higher specific Fear reported more symptoms of Anxiety.

Altogether, the findings from the EC and NE domains are consistent with Snyder et al.'s claim that their construct validity analyses should generalize to other samples, and reinforce their statement that their derived EC and NE factors "revealed specific, theoretically predicted, and meaningful patterns of links with [the] outcome measures" (p. 1145).

Positive emotionality—In contrast to EC and NE, there was little to replicate in the PE domain. Snyder et al. (2015) found mostly null concurrent associations between PE and adolescent functioning and the present findings revealed a similar pattern of null effects for the two PE factors (i.e., Surgency, Affiliation), both concurrently and prospectively. Therefore, we should consider why PE, as measured by the EATQ-R, does not seem to be related to adolescent functioning in either Snyder et al. or the present study. As Snyder et al. allude to throughout their article, the lack of associations may not reflect a problem with the PE scale, but rather with the outcomes chosen to assess its construct validity. Snyder et al. avoid making clear predictions for associations between PE and adolescent functioning measures, though they do posit that higher levels of PE (conceptualized as Surgency) should be associated reward-oriented tendencies such as ADHD and conduct problems. Indeed, although none of the associations between the PE factors and measures of adolescent functioning were significant in the present study, the effect sizes were largest for ADHD. To explore whether PE is meaningfully related to other measures of adolescent functioning, future research should test whether PE is related to more theoretically relevant measures, such as substance use (Depue, Luciana, Arbisi, Collins, & Leon, 1994) or romantic relationship quality (Robins, Caspi, & Moffitt, 2002).

Broader Implications and Recommendations

Overall, the results of the present study suggest that we have yet to discover a generalizable latent structure of temperament as measured by the EATQ-R. These findings echo previous discussions about Rothbart's theory-driven approach to developing temperament measures

(Kim, Brody, & Murry, 2003; Kotelnikova et al., 2016; Kotelnikova et al., 2017; Latham et al., 2020; Muris & Meesters, 2009) and claims from Hopwood and Donnellan (2010), who demonstrated that fitting CFAs to well-validated personality measures often results in poor model fit according to conventional standards (in part because personality researchers typically fit CFAs to items rather than composite scores). In fact, the fit indices obtained in the present study for EC, NE, and PE were similar to, or exceeded, those reported by Hopwood and Donnellan for several other widely used personality measures. Thus, the EATQ-R is not unique in exhibiting relatively poor model fit. Further, our findings also highlight how ML estimation might be problematic given the measurement properties of item-level personality data, so using WLSMV estimation or item parcels can result in better model fit.

Despite the absence of a clear structure – at the domain level and at the level of the overall questionnaire – the replicability of most concurrent associations between temperament domains and adolescent functioning suggest that temperament, as assessed via the EATQ-R, is consistently related, both concurrently and prospectively, to theoretically relevant adolescent outcomes. Thus, the present study suggests that the EATQ-R has a rather poor factor structure (both within and across temperament domains) but reasonably good validity, at least for EC and NE.

These results also highlight the quagmire that occurs when measures of temperament such as the EATQ-R are not scored consistently across studies, which results in it being “difficult to compare the results and build a systematic, replicable knowledge base” (Snyder et al., 2015, p. 1134). However, it is equally important to ensure that researchers closely match the scoring procedure with their conceptualization of the construct being assessed. Indeed, the lack of a clear structure actually empowers researchers to more carefully consider theoretical issues when determining how they use the EATQ-R. Below, we delve into each of the three temperament domains assessed by the EATQ-R and provide recommendations for scoring these domains based on findings from Snyder et al., the present study, and theories about adolescent temperament.

Effortful control—The construct of effortful control lies within a larger nomological network of self-regulatory traits including self-control, executive function, and conscientiousness (Carver, 2005). Rothbart and her colleagues have clearly articulated their theoretical conceptualization of EC in numerous publications, and there is little question in our minds that the EATQ-R EC subscales (Inhibitory Control, Activation Control, and Attention) and item content very closely map onto this theoretical conceptualization. Thus, the EC scale has strong content validity, despite its lack of a coherent factor structure. The EC scale also shows strong construct validity, as evidenced by its generalizable associations with theoretically relevant measures of adolescent functioning (i.e., lower levels of depression and ADHD, less antisocial interpersonal functioning, higher grades, less school discipline). Based on these theoretical considerations and empirical findings, we believe that the EC domain should be scored as a composite of Inhibitory Control, Activation Control, and Attention. These facets are all necessary components of the EC construct, as conceptualized by Rothbart, and a superordinate EC construct defined by all three components constitutes a theoretically, even if not empirically, coherent construct.

Notably, however, Snyder et al. found that the Activation-specific and common EC factors showed somewhat different concurrent associations with measures of adolescent functioning. The authors note that this pattern illustrates one way that EC parallels executive function, given that "there are both specific EF abilities and a common EF ability, which spans these components" (Snyder et al., 2015, p. 1143). The divergent findings for Activation-specific and common EC factors is consistent with previous research with the present sample showing that Activation Control has a different developmental trajectory across adolescence than overall EC and the Inhibitory Control and Attention Control facets (Atherton, Lawson, & Robins, 2020). On the other hand, in the present study, we did not replicate the discriminant concurrent associations reported by Snyder et al. and instead found that the common EC and Activation-specific factors showed similar associations with measures of adolescent functioning, with common EC associations tending to be stronger than Activation-specific associations. Therefore, given this discrepancy, researchers using the EATQ-R should always examine the degree to which findings observed at the EC domain level replicate across the three facets, especially for Activation Control.

Negative emotionality—Negative emotionality, as assessed by the EATQ-R, is an empirically coherent construct where both the bifactor model and the vast majority of associations with adolescent functioning generalized from Snyder et al. to the present study. Therefore, evidence suggests that researchers can use Snyder et al.'s proposed scoring method and examine both a common NE core as well as specific scores for Fear, Frustration, Shyness, Aggression, and Depressed Mood. Further, because a correlated factors model also fit the NE scale well, there is evidence that conceptualizing NE as multiple interrelated facets is also appropriate.

However, one concern is that a composite of all five facets represents an overly broad mapping of the NE domain. Although many conceptualizations of NE encompass emotions like anger and hostility that can contribute to aggression, actual aggressive *behaviors* are generally not considered part of the NE construct. Thus, a case could be made for excluding the Aggression scale from the NE domain.[14] Similarly, Neuroticism – the Big Five domain most closely aligned with NE – is conceptualized as including both high arousal (e.g., Fear, Frustration) and low arousal (e.g., Depressed Mood) negative affect, and Shyness is not a good marker of Neuroticism because it is an interstitial trait that includes aspects of low Extraversion in addition to high Neuroticism. Thus, if a researcher wants to assess the NE domain as conceptualized within a Big Five framework, it would be best to form a composite based on the Fear, Frustration, and Depressed Mood subscales of the EATQ-R, and exclude the Aggression and Shyness subscales. This seems consistent with Rothbart's conceptualization of NE, as akin to "neuroticism, that is, a general tendency to experience and express negative emotions" (Rothbart & Ahadi, 1994, p. 57). Finally, the construct of Negative Affectivity (Watson, Clark, & Tellegen, 1988) represents an even narrower conceptualization of the NE domain that would exclude Depressed Mood and include only the Fear and Frustration scales (i.e., only high arousal negative affect). Therefore, when using the EATQ-R, researchers should not only examine both common/domain and specific/

[14]Indeed, Rothbart did not include Aggression as part of the EATQ-R NE domain (Ellis & Rothbart, 2001).

facet scores of NE, they should also be clear about their conceptualization of NE and choose only the facets that map onto this construct.

Positive emotionality—As we (and Snyder et al.) have discussed, positive emotionality, as assessed via the EATQ-R, is not empirically coherent. Luckily, this fragmentation allows for fairly simple scoring guidance. In particular, researchers should never create a common PE factor from the various EATQ-R subscales, but instead should always separately score Surgency and Affiliation (along with Perceptual Sensitivity and Pleasure Sensitivity, if assessed). This is consistent with Snyder et al.'s recommendation that, "If surgency is the construct of interest, only the Surgency subscale should be used" (Snyder et al., 2015, p. 1144), as well as with Rothbart's (Rothbart & Ahadi, 1994) and Watson et al.'s (1988) theoretical conceptualization of PE. This recommendation also makes sense from a Big Five perspective because PE is most closely aligned with Big Five Extraversion, which is also sometimes labeled Surgency. In contrast, the EATQ-R Affiliation scale (which assesses the "desire for warmth and closeness with others, independent of shyness or extraversion) is most closely related to the Agreeableness domain, and neither the Perceptual Sensitivity scale (which assesses "Detection or perceptual awareness of slight, low-intensity stimulation in the environment") nor the Pleasure Sensitivity scale ("Amount of pleasure related to activities or stimuli involving low intensity, rate, complexity, novelty, and incongruity") have a close conceptual connection to any Big Five domain, although they may be loosely connected to Openness to Experience (both scales) and Neuroticism (Perceptual Sensitivity only). Consequently, it does not make sense from a Big Five perspective to ever form a composite of the Affiliation, Perceptual Sensitivity, and Pleasure Sensitivity scales, because these scales are conceptually associated with distinct Big Five domains. Thus, we recommend that researchers should use the Surgency scale to assess the superordinate temperament domain of Positive Emotionality, and use the other scales only as measures of their respective specific constructs.

Limitations and Future Directions

The present study has several limitations. First, and most notably, the EATQ-R data in the present study do not include the Perceptual Sensitivity and Pleasure Sensitivity scales that were included in Snyder et al. (2015). Consequently, we were not able to directly test the generalizability of Snyder et al.'s derived factor structure of the PE domain and instead examine a modified factor structure. It is worth noting, however, that the PE subscales did not form a coherent general factor in Snyder et al. (2015), and had very few concurrent associations with theoretically relevant variables. Second, participants in the present study differ from those in the original study not only because of ethnicity, but also because of their geographic location, immigrant status, and SES (as well as other unmeasured variables). Therefore, the suboptimal bifactor model fit results might be due to the fact that the samples in the original and present study differ in important ways, and we do not have enough information to pinpoint the driving factor responsible for discrepancies in the findings. Third, Snyder et al. (2015) used self-reports with Likert-type scales to assess depression, anxiety, and ADHD, whereas the present study used diagnostic symptom counts derived from a structured psychiatric interview. As is typically found for symptom counts, there was less variance in the depression, anxiety, and ADHD scores in the

present study compared to the original study. Nonetheless, we replicated 5/6 (83%) of the significant concurrent associations between depression, anxiety, and ADHD symptoms and latent temperament scores and we also found 19 additional significant associations between these psychopathology measures and temperament as assessed via the EATQ-R. Despite these limitations, the present study helps to calibrate our confidence in the findings from Snyder et al.; in particular, the degree to which basic dimensions of temperament have both concurrent and prospective associations with various measures of psychological functioning.

Concluding Remarks

Moving forward, researchers should continue to evaluate the structure and construct validity of the EATQ-R and examine generalizability across more diverse samples. Construct validation is a never-ending iterative process, especially when an existing scale is being used in a new context or population (Flake et al., 2017). Indeed, ongoing construct validation is especially important in research using data from participants traditionally underrepresented in psychological research, and Hernández and colleagues highlight its importance as "one of three essential components of cultural validation…" (Hernández, Nguyen, Casanova, Suárez-Orozco, & Saetermoe, 2013, p. 49). Thoughtful measurement work will also help to lay a better foundation for future applied research on youth temperament, especially with ethnic-minority samples. One method for continuing this process, as demonstrated by the present study, is taking advantage of existing longitudinal data to conduct replication-focused research. Because replications typically entail the exact analyses conducted in the original study, problems with selective reporting and *p*-hacking are minimized with replications conducted using existing data. Similarly, the hypotheses for the replication are based on findings from the original study, reducing concerns about hypothesizing after the results are known (i.e., HARKing). Further, although direct or exact replications are a vital step toward improving psychological research, the present study demonstrates that replication studies should also work to improve on previous research and not solely replicate what was done. This is especially true when there are conceptual or methodological limitations with the original research, as there often are, and when new concerns (e.g., about the limits of bifactor models) have arisen since publication of the original article. In particular, we recommend that replication studies always repeat the exact analyses conducted in the original study, which is necessary for robust examination of the replicability and generalizability of the findings, but also conduct additional (often exploratory) analyses to redress the limitations of the original study and explore alternative explanations for the findings. Together, an exact replication combined with a thoughtful extension aimed at ameliorating problems with the existing research will provide a more nuanced understanding of the replicability and generalizability of the findings.

Snyder and colleagues (2015) took an important step towards heeding Tackett and Durbin's (2017) call to focus on measurement and construct validity in youth temperament research. The present study follows in their footsteps to contribute to a more robust, generalizable science of adolescent temperament. However, the "tedious journey to construct validity" continues onward.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This research was supported by a grant from the National Institute on Drug Abuse and the National Institute on Alcohol Abuse and Alcoholism (DA017902) to Richard W. Robins. Additional project materials including R scripts to run all analyses, item wording for scales, and the supplemental preregistration are available on the Open Science Framework (OSF): https://osf.io/5rhjb/.

References

Asendorpf JB, Conner M, De Fruyt F, De Houwer J, Denissen JJA, Fiedler K, … Wicherts JM (2013). Recommendations for increasing replicability in psychology: Recommendations for increasing replicability. European Journal of Personality, 27(2), 108–119. 10.1002/per.1919

Atherton OE, Lawson KM, Ferrer E, & Robins RW (2020). The role of effortful control in the development of ADHD, ODD, and CD symptoms. Journal of Personality and Social Psychology.

Atherton OE, Lawson KM, & Robins RW (2020). The development of effortful from late childhood to young adulthood. Journal of Personality and Social Psychology.

Atherton OE, Schofield TJ, Sitka A, Conger RD, & Robins RW (2016). Unsupervised self-care predicts conduct problems: The moderating roles of hostile aggression and gender. Journal of Adolescence, 48, 1–10. 10.1016/j.adolescence.2016.01.001 [PubMed: 26820648]

Atherton OE, Tackett JL, Ferrer E, & Robins RW (2017). Bidirectional pathways between relational aggression and temperament from late childhood to adolescence. Journal of Research in Personality, 67, 75–84. 10.1016/j.jrp.2016.04.005 [PubMed: 28943676]

Atherton OE, Zheng LR, Bleidorn W, & Robins RW (2019). The co-development of effortful control and school behavioral problems. Journal of Personality and Social Psychology, 117, 659–673. [PubMed: 30035568]

Aizpitarte A, Atherton OE, & Robins RW (2017). The co-development of relational aggression and disruptive behavior symptoms from late childhood through adolescence. Clinical Psychological Science, 5, 866–873. 10.1177/2167702617708231 [PubMed: 29057169]

Bollen KA (2002). Latent variables in psychology and the social sciences. Annual Review of Psychology, 53, 605–634.

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

Brandes CM, Herzhoff K, Smack AJ, & Tackett JL (in press). The p factor and the n factor: Associations between the general factors of psychopathology and neuroticism in children. Clinical Psychological Science.

Bravo M, Woodbury-Farina M, Canino GJ, & Rubio-Stipec M (1993). The Spanish translation and cultural adaptation of the Diagnostic Interview Schedule for Children (DISC) in Puerto Rico. Culture, Medicine, & Psychiatry, 17, 329–344.

Capaldi DM & Rothbart MK (1992). Development and validation of an early adolescent temperament measure. Journal of Early Adolescence, 12, 153–173.

Carver CS (2005). Impulse and constraint: Perspectives from personality psychology, convergence with theory in other areas, and potential for integration. Personality and Social Psychology Review, 9(4), 312–333. 10.1207/s15327957pspr0904_2 [PubMed: 16223354]

Caspi A, Houts RM, Belsky DW, Goldman-Mellor SJ, Harrington H, Israel S, … Moffitt TE (2014). The p Factor: One general psychopathology factor in the structure of psychiatric disorders? Clinical Psychological Science, 2(2), 119–137. 10.1177/2167702613497473 [PubMed: 25360393]

Chen FF, West S, & Sousa K (2006). A comparison of bifactor and second-order odels of quality of life. Multivariate Behavioral Research, 41(2), 189–225. 10.1207/s15327906mbr4102_5 [PubMed: 26782910]

Cheung GW, & Rensvold RB (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. Structural Equation Modeling: A Multidisciplinary Journal, 9(2), 233–255. 10.1207/S15328007SEM0902_5

Clark DA, Donnellan MB, Robins RW, & Conger RD (2015). Early adolescent temperament, parental monitoring, and substance use in Mexican-origin adolescents. Journal of Adolescence, 41, 121–130. 10.1016/j.adolescence.2015.02.010 [PubMed: 25841175]

Clark DA, Donnellan MB, Durbin CE, Nuttall AK, Hicks BM, & Robins RW (2020). Sex, drugs, and early emerging risk: Examining the association between sexual debut and substance use across adolescence. PLOS ONE, 15(2). doi: 10.1371/journal.pone.0228432

Coplan RJ, & Bullock A (2012). Temperament and peer relationships. In Zentner M & Shiner RL (Eds.), Handbook of temperament (p. 442–461). The Guilford Press.

Costello EJ, Edelbrock CS, & Costello AJ (1985). Validity of the NIMH diagnostic interview schedule for children: A comparison between psychiatric and pediatric referrals. Journal of Abnormal Child Psychology, 13, 579–595. 10.1007/BF00923143 [PubMed: 4078188]

De Bolle M, & Tackett JL (2013). Anchoring bullying and victimization in children within a five-factor model-based person-centred framework: Bullying, victimization and personality: a person-centered approach. European Journal of Personality, 27(3), 280–289. 10.1002/per.1901

Depue RA, Luciana M, Arbisi P, Collins P, & Leon A (1994). Dopamine and the structure of personality: Relation of agonist-induced dopamine activity to positive emotionality. Journal of Personality and Social Psychology, 67, 485–498. 10.1037/0022-3514.67.3.485 [PubMed: 7965602]

Dueber DM (2017). Bifactor Indices Calculator: A Microsoft Excel-based tool to calculate various indices relevant to bifactor CFA models. 10.13023/edp.tool.01 [Also available at http://sites.education.uky.edu/apslab/resources/]

Eisenberg N, Ma Y, Chang L, Zhou Q, West SG, & Aiken L (2007). Relations of effortful control, reactive undercontrol, and anger to Chinese children's adjustment. Development and Psychopathology, 19(02). 10.1017/S0954579407070198

Eisenberg N, Valiente C, Spinrad TL, Cumberland A, Liew J, Reiser M, Zhou Q, & Losoya SH (2009). Longitudinal relations of children's effortful control, impulsivity, and negative emotionality to their externalizing, internalizing, and co-occurring behavior problems. Developmental Psychology, 45(4), 988–1008. 10.1037/a0016213 [PubMed: 19586175]

Ellis LK, & Rothbart MK (2001). Revision of the Early Adolescent Temperament Questionnaire. Poster presented at the Biennial Meeting of the Society for Research in Child Development in Minneapolis, Minnesota.

Else-Quest NM, Hyde JS, Goldsmith HH, & Van Hulle CA (2006). Gender differences in temperament: A meta-analysis. Psychological Bulletin, 132(1), 33–72. 10.1037/0033-2909.132.1.33 [PubMed: 16435957]

Finkel EJ, Eastwick PW, & Reis HT (2017). Replicability and other features of a high-quality science: Toward a balanced and empirical approach. Journal of Personality and Social Psychology, 113(2), 244–253. 10.1037/pspi0000075 [PubMed: 28714730]

Flake JK, Pek J, & Hehman E (2017). Construct validation in social and personality research: Current practice and recommendations. Social Psychological and Personality Science, 8, 370–378. 10.1177/1948550617693063

González Salinas MDC, Hidalgo Montesinos MD, Carranza Carnicero JA and Ato Garcia M (2000). Preparation of a Spanish adaptation of the Infant Behavior Questionnaire for the measurement of temperament in childhood. Psicothema, 12, 513–519.

González Salinas MDC, Carranza Carnicero JA and Hidalgo Montesinos MD (1999). Adaptation to the Spanish population of the "Toddler Behavior Assessment Questionnaire" to measure temperament in childhood. Behavioral Sciences Methodology, 1, 207–218.

Hernández MG, Nguyen J, Casanova S, Suárez-Orozco C, & Saetermoe CL (2013). Doing no harm and getting it right: Guidelines for ethical research with immigrant communities. New Directions for Child and Adolescent Development, 2013(141), 43–60. 10.1002/cad.20042 [PubMed: 24038806]

Hoffmann M, Pérez JC, García C, Rojas G, & Martínez V (2017). Chilean adaptation and validation of the Early Adolescent Temperament Questionnaire-Revised Version. Frontiers in Psychology, 8, 2131. 10.3389/fpsyg.2017.02131 [PubMed: 29326616]

Hopwood CJ, & Donnellan MB (2010). How should the internal structure of personality inventories be evaluated? Personality and Social Psychology Review, 14(3), 332–346. 10.1177/1088868310361240 [PubMed: 20435808]

Hu L, & Bentler PM (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. Psychological Methods, 3, 424–453.

Kim S, Brody GH, & Murry VM (2003). Factor structure of the Early Adolescent Temperament Questionnaire and measurement invariance across gender. The Journal of Early Adolescence, 23(3), 268–294. 10.1177/0272431603254178

Kotelnikova Y, Olino TM, Klein DN, Kryski KR, & Hayden EP (2016). Higher- and lower-order factor analyses of the Children's Behavior Questionnaire in early and middle childhood. Psychological Assessment, 28(1), 92–108. 10.1037/pas0000153 [PubMed: 26029946]

Kotelnikova Y, Olino TM, Klein DN, Mackrell SVM, & Hayden EP (2017). Higher and lower order factor analyses of the Temperament in Middle Childhood Questionnaire. Assessment, 24(8), 1050–1061. 10.1177/1073191116639376 [PubMed: 27002124]

Kovacs M (1985). The Children's Depression, Inventory (CDI). Psychopharmacology Bulletin, 21, 995–998. [PubMed: 4089116]

Latham MD, Dudgeon P, Yap MBH, Simmons JG, Byrne ML, Schwartz OS, Ivie E, Whittle S, & Allen NB (2020). Factor structure of the Early Adolescent Temperament Questionnaire–Revised. Assessment, 27(7), 1547–1561. 10.1177/1073191119831789 [PubMed: 30788984]

Loevinger J (1957). Objective tests as instruments of psychological theory. Psychological Reports, 3, 635–694.

March JS, Parker JD, Sullivan K, Stallings P, & Conners CK (1997). The Multidimensional Anxiety Scale for Children (MASC): Factor structure, reliability, and validity. Journal of the American Academy of Child & Adolescent Psychiatry, 36, 554–565. 10.1097/00004583-199704000-00019 [PubMed: 9100431]

Marsh HW, Hau K-T, & Wen Z (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. Structural Equation Modeling: A Multidisciplinary Journal, 11(3), 320–341. 10.1207/s15328007sem1103_2

Meisenberg G, & Williams A (2008). Are acquiescent and extreme response styles related to low intelligence and education? Personality and Individual Differences, 44(7), 1539–1550. 10.1016/j.paid.2008.01.010

Morgan G, Hodge K, Wells K, & Watkins M (2015). Are fit indices biased in favor of bifactor models in cognitive ability research?: A Comparison of fit in correlated factors, higher-order, and bi-factor models via Monte Carlo simulations. Journal of Intelligence, 3(1), 2–20. 10.3390/jintelligence3010002

Muris P, & Meesters C (2009). Reactive and Regulative Temperament in Youths: Psychometric Evaluation of the Early Adolescent Temperament Questionnaire-Revised. Journal of Psychopathology and Behavioral Assessment, 31(1), 7–19. 10.1007/s10862-008-9089-x

Muris P, Meesters C, & Blijlevens P (2007). Self-reported reactive and regulative temperament in early adolescence: Relations to internalizing and externalizing problem behavior and "Big Three" personality factors. Journal of Adolescence, 30(6), 1035–1049. 10.1016/j.adolescence.2007.03.003 [PubMed: 17467051]

Murray AL, & Johnson W (2013). The limitations of model fit in comparing the bi-factor versus higher-order models of human cognitive ability structure. Intelligence, 41(5), 407–422.

Nigg JT (2006). Temperament and developmental psychopathology. Journal of Child Psychology and Psychiatry, 47(3–4), 395–422. 10.1111/j.1469-7610.2006.01612.x [PubMed: 16492265]

Ojanen T, Findley D, & Fuller S (2012). Physical and relational aggression in early adolescence: Associations with narcissism, temperament, and social goals. Aggressive Behavior, 38, 99–107. [PubMed: 22331610]

Prinstein MJ, Boergers J, & Vernberg EM (2001). Overt and relational aggression in adolescents: Social-psychological adjustment of aggressors and victims. Journal of Clinical Child & Adolescent Psychology, 30(4), 479–491. 10.1207/S15374424JCCP3004_05

Putnick DL, & Bornstein MH (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. Developmental Review, 41, 71–90. 10.1016/j.dr.2016.06.004 [PubMed: 27942093]

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Reise SP, Moore TM, & Haviland MG (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. Journal of Personality Assessment, 92(6), 544–559. 10.1080/00223891.2010.496477 [PubMed: 20954056]

Revelle W (2018) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 1.8.12.

Robins RW, Caspi A, & Moffitt T (2002). It's not just who you're with, it's who you are: Personality and relationship experiences across multiple relationships. Journal of Personality, 70, 925–964. [PubMed: 12498360]

Robins RW, Donnellan MB, Widaman KF, & Conger RD (2010). Evaluating the link between self-esteem and temperament in Mexican origin early adolescents. Journal of Adolescence, 33(3), 403–410. 10.1016/j.adolescence.2009.07.009 [PubMed: 19740537]

Rodriguez A, Reise SP, & Haviland MG (2015). Evaluating bifactor models: Calculating and interpreting statistical indices. Psychological Methods. 10.1037/met0000045

Rosseel Y (2012). lavaan: An R package for structural equation modeling. Journal of Statistical Software, 48(2), 1–36. URL http://www.jstatsoft.org/v48/i02/

Rothbart MK (2007). Temperament, development, and personality. Current Directions in Psychological Science, 16, 207–212. 10.1111/j.1467-8721.2007.00505.x

Rothbart MK (2011). Becoming who we are: Temperament and personality in development. New York, NY: Guilford Press.

Rothbart MK, & Ahadi SA (1994). Temperament and the development of personality. Journal of Abnormal Psychology, 103(1), 55–66. [PubMed: 8040481]

Rothbart MK, Ahadi SA, & Hershey KL (2020). Temperament and social behavior in childhood. Merrill-Palmer Quarterly, 40, 21–39.

Rothbart MK, & Bates JE (2006). Temperament. In Eisenberg N, Damon W, & Lerner RM (Eds.), Handbook of child psychology: Social, emotional, and personality development (pp. 99–166). Hoboken, NJ, US: John Wiley & Sons Inc.

Rothbart MK, Ahadi SA, & Evans DE (2000). Temperament and personality: Origins and outcomes. Journal of Personality and Social Psychology, 78(1), 122–135. [PubMed: 10653510]

Schwab-Stone M, Shaffer D, Dulcan M, Jensen P, Fisher P, Bird H, …Rubio-Stipec, & Rae D (1996). Criterion validity of the NIMH Diagnostic Interview Schedule for Children Version 2.3 (DISC 2.3). Journal of the American Academy of Child & Adolescent Psychiatry, 35, 878–888. 10.1097/00004583-199607000-00013 [PubMed: 8768347]

Snyder HR, Gulley LD, Bijttebier P, Hartman CA, Oldehinkel AJ, Mezulis A, … Hankin BL (2015). Adolescent emotionality and effortful control: Core latent constructs and links to psychopathology and functioning. Journal of Personality and Social Psychology, 109, 1132–1149. [PubMed: 26011660]

Swanson JM, Kraemer HC, Hinshaw SP, Arnold LE, Conners CK, Abikoff HB, ... Wu M (2001). Clinical relevance of the primary findings of the MTA: Success rates based on severity of ADHD and ODD symptoms at the end of treatment. Journal of the American Academic of Child & Adolescent Psychiatry, 40, 168–179. 10.1097/00004583-200102000-00011

Tackett JL, & Durbin CE (2017). Advances in research on youth personality: Introduction to the special issue. Journal of Research in Personality, 67, 1–2.

Tackett JL, Kushner SC, Herzhoff K, Smack AJ, & Reardon KW (2014). Viewing relational aggression through multiple lenses: Temperament, personality, and personality pathology. Development and Psychopathology, 26(3), 863–877. 10.1017/S0954579414000443 [PubMed: 25047304]

Taylor ZE, Widaman KF, & Robins RW (2018). Longitudinal relations of economic hardship and effortful control to active coping in Latino youth. Journal of Research on Adolescence, 28, 396–411. 10.1111/jora.12338 [PubMed: 28851024]

Valiente C, Eisenberg N, Spinrad TL, Haugen R, Thompson MS, & Kupfer A (2013). Effortful control and impulsivity as concurrent and longitudinal predictors of academic achievement. The Journal of Early Adolescence, 33(7), 946–972. 10.1177/0272431613477239

Van De Schoot R, Schmidt P, De Beuckelaer A, Lek K, & Zondervan-Zwijnenburg M (2015). Editorial: Measurement Invariance. Frontiers in Psychology, 6. 10.3389/fpsyg.2015.01064

Viñas F, González M, Gras E, Jane C, & Casas F (2015). Psychometric properties of the EATQ-R among a sample of Catalan-speaking Spanish adolescents. Universitas Psychologica, 14, 747–758.

Watson D, Clark LA, & Carey G (1988). Positive and negative affectivity and their relation to anxiety and depressive disorders. Journal of Abnormal Psychology, 97, 346–353. [PubMed: 3192830]

Watson D, Clark LA, & 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. [PubMed: 3397865]

Watts AL, Poore HE, & Waldman ID (2019). Riskier tests of the validity of the bifactor model of psychopathology. Clinical Psychological Science, 1–19.

Table 1

Comparison of Sample Characteristics in Snyder et al. (2015) versus the Present Study

	Snyder et al. (2015)		Present study
	Factor Structure	**Concurrent Validity**	**All Analyses**
N	1013	562	586
Age	13.0 (2.57)	13.6 (2.36)	12.8 (0.49)
Gender (% female)	55%	55%	50%
Race/Ethnicity	Primarily White	Primarily (69%) White	Latinx (Mexican-origin)
Nationality	American, Dutch, Belgian	American	American
Geographic location	WA, CO, NJ, The Netherlands, Belgium	CO, NJ	Northern CA
Household income	No information	Median = $86,500	Median = $32,500
Parent Education Level	No information	No information	Median = 9th grade

Note. Age is presented in the following format: Mean (*SD*). WA = Washington, CO = Colorado, NJ = New Jersey, CA = California.

Table 2

Replication of Snyder et al. Bifactor Model Fit Statistics

Model	χ^2 (*df*)	χ^2/df	CFI	RMSEA
Effortful Control	411.65(81)* [414.10(81)*]	5.08 [5.11]	.72 [.84]	.084 [.066]
Negative Emotionality	853.57(343)* [1064.67(343)*]	2.49 [3.10]	.87 [.89]	.051 [.047]
Positive Emotionality	108.08(25)* [380.32(116)*]	4.32 [3.28]	.86 [.91]	.076 [.049]
Full Model	3243.15(1252)* [4693.71(1731)*]	2.59 [2.71]	.73 [.79]	.052 [.043]

Note. CFI = confirmatory fit index; RMSEA = root mean square error of approximation. Values in brackets are fit statistics from the hold-out sample in Snyder et al. (2015) for comparison.

* $p < .001$.

Table 3

EATQ-R Full Model Estimated Factor Intercorrelations

	Common EC	Activation-specific	Surgency	Affiliation
Common NE	−.09 [−.36*]	−.72* [.30*]	.18 [−.25*]	.55*
Aggression-specific	−.36* [−.43*]	−.16 [−.17]	.04 [.17*]	−.27*
Depressed mood-specific	−.73* [−.40*]	.35 [.09]	−.03 [−.22*]	−.77*
Fear-specific	.17 [−.17]	.26 [.29]	−.18 [−.48*]	−.09
Frustration-specific	−.20 [−.41*]	−.22 [−.13]	−.20 [.17*]	−.26
Shyness-specific	−.13 [−.10*]	−.02 [−.08]	−.09 [−.21*]	−.33*
Common EC			.04 [.11]	.66*
Activation-specific			−.21 [−.26*]	−.62*
Surgency	.04 [.11]	−.21 [−.26*]		

Note. Blanks in the table indicate auto-correlations or factors constrained not to correlate (e.g., specific factors within each domain do not correlate with each other or with their Common factor because their shared variance is already captured by their Common factor). Values in brackets are coefficients from the hold-out dataset in Snyder et al. (2015) for comparison.

* $p < .0005$.

Table 4

Modified Bifactor Model Fit Statistics

Model	χ^2 (*df*)	χ^2/df	CFI	RMSEA
Effortful Control				
Original (Snyder et al.) model	411.65(81)*	5.08	.72	.084
No correlated residuals	446.71(85)*	5.26	.70	.086
Modified correlated residuals	368.68(81)*	4.55	.76	.078
Random correlated residuals	423.41(81)*	5.23	.71	.085
Negative Emotionality				
Original (Snyder et al.) model	853.57(343)*	2.49	.87	.051
No correlated residuals	906.04(349)*	2.60	.86	.053
Modified correlated residuals	785.71(343)*	2.29	.89	.047
Random correlated residuals	897.46(343)*	2.62	.86	.053
Positive Emotionality				
Original (Snyder et al.) model	108.08(25)*	4.32	.86	.076
No correlated residuals	135.42(26)*	5.21	.82	.085
Modified correlated residuals	108.08(25)*	4.32	.86	.076
Random correlated residuals	126.85(25)*	5.07	.83	.084

Note. CFI = confirmatory fit index; RMSEA = root mean square error of approximation. “Original (Snyder et al.) model” refers to the fit of models retained in Snyder et al. (2015) when evaluated using data from the present study. “No correlated residuals” refers to the models with no correlated residuals. “Modified correlated residuals” refers to the models with adjusted correlated residuals based on the largest modification indices. “Random correlated residuals” refers to models with random pairs of correlated residuals.

* $p < .001$.

Table 5

New Model Comparison Fit Statistics

Temperament Domain	Model	$\chi^2(df)$	χ^2/df	CFI	RMSEA	AIC	BIC	Sample-size adjusted BIC
Effortful Control	Single factor	580.20(104)	5.58	.65	.089	22757	22967	22814
	Correlated factors	562.81(102)	5.52	.66	.088	22744	22962	22803
	Bifactor model	400.67(88)	4.55	.77	.078	22610	22889	22686
	Hierarchical model	--	--	--	--	--	--	--
Negative Emotionality	Single factor	2190.17(377)	5.81	.54	.091	39658	40037	39761
	Correlated factors	923.20(367)	2.52	.86	.051	38411	38834	38526
	Bifactor model	911.53(348)	2.62	.86	.053	38437	38943	38575
	Hierarchical model	1071.38(372)	2.88	.82	.057	38549	38950	38658
Positive Emotionality	Single factor	276.01(44)	6.27	.64	.096	16710	16854	16749
	Correlated factors	189.89(43)	4.42	.77	.077	16626	16774	16666
	Bifactor model	73.69(33)	2.23	.94	.046	16529	16721	16582
	Hierarchical model	189.89(42)	4.52	.77	.078	16628	16780	16669

Note. CFI = confirmatory fit index; RMSEA = root mean square error of approximation. AIC = Akaike Information Criteria. BIC = Bayesian Information Criteria. Hierarchical EC model did not converge.

Table 6

Descriptive Statistics for Measures of Adolescent Functioning

	Concurrent Assessment (Age 13)			Predictive Assessment (Age 15)		
	Mean	**SD**	**Range**	**Mean**	**SD**	**Range**
Depression	3.99	3.83	0 – 18	4.08	4.22	0–18
Anxiety	2.37	1.77	0–11	1.96	1.84	0–10.5
ADHD – Total	2.02	2.47	0–14	2.03	2.47	0–15.5
ADHD – Inattention	1.01	1.32	0–8	1.07	1.42	0–8
ADHD – Hyperactivity	1.01	1.49	0–10	.97	1.41	0–8
Antisocial Interpersonal Functioning	1.08	.18	1–4	1.08	.18	1–4
Victim Interpersonal Functioning	1.14	.26	1–4	1.09	.20	1–4
School Grades	4.18	.91	1–5	3.95	1.08	1–5
School Discipline	1.23	.58	1–4	1.25	.60	1–4

J Pers Soc Psychol. Author manuscript; available in PMC 2022 November 01.

Table 7

Bifactor Models: Concurrent and Prospective Correlations between EATQ-R Factors and Measures of Adolescent Functioning (Compared with Snyder et al. Findings)

	Effortful Control	
	Common EC	Activation specific
Depression	**−.33** *[−.16 *] /*−.58**	**−.16** *[−.11] /.06*
Anxiety	**−.13** [−.05] /*−.38**	**−.05** [−.04] /*−.17*
ADHD – Total	**−.39** *[−.20 *] /*−.25**	**−.20** *[−.15] /.23*
ADHD - Inattention	**−.42** *[−.24 *] /*−.21*	**−.18** *[−.12] /*−.07*
ADHD – Hyperactivity	**−.26** *[−.10] /.15*	**−.17** *[−.13] /.05*
Antisocial Interpersonal Functioning	**−.22** *[−.18 *] /*−.45**	**−.14** [−.09] /*−.04*
Victim Interpersonal Functioning	**−.11** [−.14] /*−.35**	**−.06** [.04] /*.07*
School Grades	**.31** *[.23 *] /.36**	**.03** [.01] /*−.06*
School Discipline	**−.18** *[−.22 *] /*−.18**	**−.01** [−.03] /*−.03*

	Negative Emotionality					
	Common NE	Aggression spec.	Depression spec.	Fear spec.	Frustration spec.	Shyness spec.
Depression	**.38** *[.26 *]/.57**	**.12** [.04] / *.12*	**.27** *[.20 *] / .50**	**−.06** [−.05] / *−.14*	**.16** *[.04] / *−.07*	**−.04** [.02] / *−.01*
Anxiety	**.41** *[.23 *]/.75**	**−.06** [−.08] / *−.09*	**.09** [.05] / *.00*	**.18** *[.07] / *.05*	**.10** [.03] / *−.18*	**.12** [.01] / *.16*
ADHD – Total	**.30** *[.28 *]/*−.08*	**.25** *[.19 *]/ *.16*	**.19** *[.09] / *.21*	**−.14** [−.05] / *−.01*	**.23** *[.11] / *.19*	**−.10** [−.06] / *−.12*
ADHD – Inattention	**.29** *[.23 *]/*−.02*	**.22** *[.18 *]/ *.05*	**.18** *[.09] / *.19*	**−.09** [−.05] / *.10*	**.20** *[.09] / *.14*	**−.04** [−.02] / *.08*
ADHD – Hyperactivity	**.21** *[.25 *]/−.10	**.22** *[18] /*−.12*	**.12** [.01] / *.14*	**−.13** [−.04] / *.08*	**.19** *[.10] / *.07*	**−.14** [−.11] / *.10*
Antisocial Interpersonal Funct.	**.21** *[.15]/.36**	**.32** *[.32 *]/.46**	**.13** [.12] / *.02*	**−.09** [−.10] / *−.13*	**.14** [.08] / *−.16*	**−.11** [−.11] / *−.09*
Victim Interpersonal Funct.	**.25** *[.16 *]/.34**	**.08** [.06]/*.27**	**.20** *[.14] / *.19*	**.07** [.03] / *−.05*	**.07** [.02] / *−.12*	**−.03** [−.00] / *−.09*
School Grades	**−.07** [.06] / *−.11*	**−.15** *[−.17 *]/*−.35**	**−.11** [−.09] / *−.10*	**.01** [.07] / *−.25**	**.00** [.03] / *−.04*	**.05** [.05] / *−.01*
School Discipline	**.00** [.07] / *−.11*	**.19** *[.17 *]/.26 *	**.01** [.09] / *.20*	**−.12** [−.08] / *.04*	**.11** [.05] / *.20*	**−.08** [−.06] / *−.10*

	Positive Emotionality

	Surgency	Affiliation
Depression	**.07** [.08] / *−.15*	**.04** [.08]
Anxiety	**.06** [.09] / *−.14*	**.04** [.07]
ADHD – Total	**.10** [.18*] / *.12*	**.07** [.11]
ADHD - Inattention	**.08** [.15]/ *−.02*	**.03** [.05]
ADHD - Hyperactivity	**.09** [.12] / *.01*	**.09** [.12]
Antisocial Interpersonal Functioning	**.03** [.03] / *−.01*	**−.00** [.04]
Victim Interpersonal Functioning	**.04** [.01] / *.05*	**.03** [.04]
School Grades	**−.05** [.03] / *−.05*	**.11** [.19*]
School Discipline	**.03** [.01] / *.11*	**−.06** [−.07]

Note. Bolded values are concurrent associations from the present study. Values in brackets are prospective associations from the present study. Values in italics after the slash are concurrent coefficients from Snyder et al. (2015) for comparison. EATQ-R = Early Adolescent Temperament Questionnaire-Revised; EC = Effortful Control; NE = Negative Emotionality; PE = Positive Emotionality; ADHD = Attention deficit/hyperactivity disorder; Spec. = Specific; Funct. = Functioning.

*$p < .0003$

Table 8

Correlated Factors Models: Concurrent and Prospective Correlations between EATQ-R Factors and Measures of Adolescent Functioning

	Effortful Control		
	Activation Control	Attention	Inhibitory Control
Depression	−.30* [−.16*]	−.29* [−.14]	−.28* [−.12]
Anxiety	−.09 [−.06]	−.10 [−.04]	−.11 [−.03]
ADHD - Total	−.38* [−.19*]	−.36* [−.16]	−.34* [−.14]
ADHD - Inattention	−.40* [−.23*]	−.39* [−.20*]	−.38* [−.19*]
ADHD - Hyperactivity	−.26* [−.11]	−.24* [−.07]	−.23* [−.05]
Antisocial Interpersonal Functioning	−.21* [−.16*]	−.20* [−.16*]	−.20* [−.16*]
Victim Interpersonal Functioning	−.10 [−.12]	−.10 [−.14]	−.10 [−.14]
School Grades	.30* [.23*]	.30* [.23*]	.29* [.23*]
School Discipline	−.18* [−.22*]	−.18* [−.21*]	−.17* [−.20*]

	Negative Emotionality				
	Aggression	Depressed Mood	Fear	Frustration	Shyness
Depression	.32* [.20*]	.41* [.29*]	.22* [.14]	.36* [.23*]	.12 [.13]
Anxiety	.17* [.07]	.41* [.22*]	.41* [.20*]	.36* [.18*]	.30* [.11]
ADHD - Total	.40* [.31*]	.30* [.27*]	.09 [.11]	.35* [.29*]	.03 [.06]
ADHD - Inattention	.37* [.28*]	.30* [.22*]	.13 [.08]	.33* [.25*]	.07 [.07]
ADHD - Hyperactivity	.32* [.26*]	.21* [.22*]	.05 [.10]	.27* [.24*]	−.03 [.00]
Antisocial Interpersonal Funct.	.37* [.33*]	.22* [.16*]	.04 [−.01]	.26* [.18*]	−.01 [−.04]
Victim Interpersonal Funct.	.17* [.12]	.27* [.18*]	.20* [.12]	.23* [.14]	.08 [.07]
School Grades	−.16* [−.12]	−.10 [.02]	−.02 [.10]	−.07 [.03]	.01 [.07]
School Discipline	.17* [.17*]	−.02 [.07]	−.10 [−.02]	.07 [.09]	−.09 [−.03]

	Positive Emotionality	
	Surgency	Affiliation

Depression	.08 [.08]	.02 [.04]
Anxiety	.07 [.08]	.04 [.07]
ADHD - Total	.11 [.19*]	.04 [.07]
ADHD - Inattention	.09 [.15]	−.00 [.02]
ADHD - Hyperactivity	.10 [.13]	.06 [.10]
Antisocial Interpersonal Functioning	.03 [.02]	−.02 [.02]
Victim Interpersonal Functioning	.03 [.01]	.02 [.02]
School Grades	−.05 [.03]	.13 [.20*]
School Discipline	.02 [.01]	−.09 [−.08]

Note. Values are concurrent associations from the present study. Values in brackets are prospective associations from the present study. EATQ-R = Early Adolescent Temperament Questionnaire-Revised; EC = Effortful Control; NE = Negative Emotionality; PE = Positive Emotionality; ADHD = Attention deficit/hyperactivity disorder; Spec. = Specific; Funct. = Functioning.

* $p < .0005$