AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Assessment

ISSN: 1040-3590

2026, Vol. 38, No. 4, 295–306
https://doi.org/10.1037/pas0001441

How Often Is "Often"? Improving Assessment of the Externalizing Spectrum Using Absolute Frequency

Isaac T. Petersen[1], Zachary Demko[1], Philipp Doebler[2], Loreen Sabel[2], Jacob J. Oleson[3], and Robert F. Krueger[4]

[1] Department of Psychological and Brain Sciences, University of Iowa
[2] Fachgebiet für Statistische Methoden in den Sozialwissenschaften, Technische Universität Dortmund
[3] Department of Biostatistics, University of Iowa
[4] Department of Psychology, University of Minnesota

Nearly all questionnaires of externalizing problems use vague quantifiers of relative frequency (e.g., rarely/sometimes/often) or true/false statements. Vague quantifiers have many problems, including imprecision and low interpretability. An alternative is numeric quantifiers that quantify, in absolute frequency, how many times the person engaged in the behavior during a given time frame. This study evaluates whether absolute frequency provides utility for assessing the externalizing spectrum. Participants included adults recruited online and college students, for a combined sample of 1,237 adults (290 males; 947 females) spanning 18–92 years of age. A subset of items was adapted from the Externalizing Spectrum Inventory to assess absolute frequency, supplemented with additional items to ensure broad coverage. Using a 30-day reference period, participants indicated how many times they engaged in each behavior per day, per week, in the past month, or in the prior year. Externalizing problems showed age-related decreases from early to later adulthood. On average, men showed greater externalizing problems than women in early and older adulthood; women showed greater externalizing problems than men in middle adulthood. Latent scores derived from absolute frequency items demonstrated convergent validity with a widely used measure of externalizing problems (Adult Self-Report), discriminant validity with respect to internalizing problems, and criterion and incremental validity in relation to functional impairment and inhibitory control. Count data led to greater precision—less uncertainty in the estimate of each person's level of externalizing problems—than dichotomized versions of the items. Findings suggest there is key utility in assessing absolute frequency of externalizing behavior.

Public Significance Statement

The present study suggests that having respondents specify the number of times they engage in various problem behaviors is more useful than having respondents indicate how often the behavior occurred using subjective labels such as "rarely," "sometimes," and "often." Having respondents specify the number of times they engage in various problem behaviors may lead to better assessment for behavior problems.

Keywords: externalizing behavior problems, absolute frequency versus relative frequency, numeric quantifiers versus vague quantifiers, item response theory, adults

Supplemental materials: https://doi.org/10.1037/pas0001441.supp

This article was published Online First December 4, 2025.
Jaime L. Anderson served as action editor.
Isaac T. Petersen https://orcid.org/0000-0003-3072-6673
Zachary Demko https://orcid.org/0000-0002-5717-6205
Philipp Doebler https://orcid.org/0000-0002-2946-8526
Loreen Sabel https://orcid.org/0000-0002-9832-8842
Jacob J. Oleson https://orcid.org/0000-0001-6343-3274
Robert F. Krueger https://orcid.org/0000-0001-9127-5509

The instruments, data, a data dictionary, analysis scripts, and a computational notebook for the present study are published at https://osf.io/49m8q (Petersen et al., 2025a). The study hypotheses were preregistered at https://osf.io/cthju (Petersen & Demko, 2024). The present study was approved by the University of Iowa Institutional Review Board (Study No.: 202305374). The authors complied with the American Psychological Association ethical standards in the treatment of participants. The authors have no conflicts of interest to disclose.

The project was supported by Grant UL1TR002537 from the National Center for Advancing Translational Sciences. Isaac T. Petersen was funded by Grant HD098235 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health. Zachary Demko was funded by Grants T32GM108540 and T32GM149386 from the National Institute of General Medical Sciences. Robert F. Krueger was supported partly by National Institute on Aging, National Institutes of Health, Grants R01AG053217, R01AG077742, and U19AG51426. The preprint of this article was posted at https://doi.org/10.31234/osf.io/r9vqz_v3 (Petersen et al., 2025b).

Isaac T. Petersen played a lead role in conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, supervision, validation, visualization,

continued

In the Woody Allen movie *Annie Hall*, there is a scene with a split screen of Annie Hall and Alvy Singer—who are in a romantic relationship. On one side, Annie talks to her psychiatrist; on the other side, Alvy talks to his psychiatrist. Alvy and Annie are each asked by their psychiatrist how often they have sex with one another. Alvy responds, "*Hardly ever*; maybe three times a week." Annie responds, "*Constantly*; I'd say three times a week." That the same absolute frequency can be expressed as different relative frequencies illustrates a key challenge with assessing behavior frequency. The present study examines whether consideration of absolute frequency is useful for assessing the externalizing spectrum—a spectrum of disinhibition and antagonism encompassing behaviors such as aggression and substance use.

Frequency—in addition to onset, timing, duration, intensity, function, and resulting impairment—is a crucial dimension to consider when assessing behavior and psychopathology. Frequency is the most common dimension assessed regarding behavior (Schaeffer, 1991) and can be assessed in terms of absolute or relative frequency. Relative frequency may indicate, for instance, that a person engaged in a given behavior (e.g., binge drinking) "often," whereas absolute frequency may indicate that they engaged in the behavior "three times in the past week." Below, we outline key differences between assessments of absolute versus relative frequency and reasons to prefer absolute frequency when assessing externalizing behavior.

Ways to Assess Behavior Frequency

There are multiple ways of assessing behavior frequency in a questionnaire. The most common ways of assessing frequency of externalizing behavior include true/false and Likert-type items. An example true/false item to assess whether a person frequently engages in externalizing behavior is: "True or False: My child often breaks the rules" (Wirt et al., 2001). Because true/false response options permit indicating endorsement or agreement rather than behavioral frequency per se, the present review focuses on Likert-type items and other formats that intrinsically permit respondents to indicate how often behaviors occur. Likert-type items commonly use vague quantifiers that assess relative frequency. Vague quantifiers are subjective labels that are intended to reflect differing frequencies of the target behavior. For instance, an example Likert-type item might ask whether the person breaks rules "never," "rarely," "sometimes," "often," or "very often." Examples of questionnaires that assess externalizing behavior using vague quantifiers are in Supplemental Table S1.

There are other ways of assessing behavior frequency, such as with numeric quantifiers that assess absolute frequency. Absolute frequency refers to the actual number of occurrences of the behavior. For instance, a person might indicate how many times in the prior month they stole from a store. To the extent that numeric and vague quantifiers each capture the same frequency information, results based on either quantifier should not differ (Marincic, 2012). However, studies indicate that findings based on vague quantifiers may differ from findings based on numeric quantifiers—in terms of different interitem associations (i.e., different factor structures; Marincic, 2012) and different associations with external criteria (e.g., Cole & Korkmaz, 2012). Below, we outline differences between vague and numeric quantifiers.

Vague Quantifiers to Assess Relative Frequency

A strength of vague quantifiers, and arguably why they are so commonly used, is that they are quick and relatively easy for respondents to rate. However, vague quantifiers have many weaknesses, which are summarized in Supplemental Table S2. Five key weaknesses are summarized here. First, they are subjective. In a standard questionnaire, there is no operational definition of "sometimes" or "often." It is therefore not known what frequency "rarely" corresponds to for a given person and behavior. Second, vague quantifiers are imprecise. The same label (e.g., "often") could span a wide range of frequency even for the same respondent. Third, most common applications of vague quantifiers uncritically assume that the quantifiers are meaningfully ordered in their degree with equal intervals. Fourth, the meaning of a quantifier can differ based on other quantifiers in the set (Bradburn & Miles, 1979) and their order (Chan, 1991). Respondents tend to choose the first response option that is acceptable to them among the ordered response options, thus leading to a primacy effect and satisficing (Chan, 1991).

Fifth, likely because of their subjectivity and imprecision, the meaning of a given vague quantifier (e.g., "often") differs by item characteristics and respondent demographic and reference groups. Differing interpretations is a crucial limitation of vague quantifiers and is described in detail below. For example, the meanings of vague quantifiers differ depending on item-level characteristics. As examples, the meaning of a given vague quantifier can differ depending on the behavior it refers to (Bradburn & Miles, 1979). For instance, prior work has shown that a given vague quantifier, such as "often," refers to a lower absolute frequency for less common events compared to more common events (Pepper & Prytulak, 1974). Thus, "often" may indicate lower frequency for externalizing behaviors with a lower base rate, such as physical aggression, versus externalizing behaviors with a higher base rate, such as arguing.

Moreover, meanings assigned to vague quantifiers differ between demographic groups. For instance, several studies have found that vague quantifiers differed in meaning as a function of the respondent's race, education, age, socioeconomic status, and geographic region (Borgers et al., 2003; Schaeffer, 1991; Wright et al., 1994). When group- or age-related differences in scores are observed, it is unclear the extent to which such differences reflect a difference in interpretation of the vague quantifier versus a true difference in construct level, confounding group-related differences in scores. Further, there are cultural differences in response styles,

writing–original draft, and writing–review and editing. Zachary Demko played a supporting role in conceptualization, data curation, investigation, and writing–review and editing and an equal role in methodology. Philipp Doebler played a supporting role in writing–review and editing and an equal role in formal analysis. Loreen Sabel played a supporting role in formal analysis and writing–review and editing. Jacob J. Oleson played a supporting role in formal analysis and writing–review and editing. Robert F. Krueger played a supporting role in methodology and writing–review and editing.

Correspondence concerning this article should be addressed to Isaac T. Petersen, Department of Psychological and Brain Sciences, University of Iowa, G60 Psychological and Brain Sciences Building, Iowa City, IA 52242, United States. Email: isaac-t-petersen@uiowa.edu

including the likelihood that individuals of different cultures select midpoint versus extreme response options (e.g., Batchelor & Miao, 2016).

Such group difference in interpretation may be due in part to individuals' reference groups for norms around a behavior. Asking questions with vague quantifiers leads respondents to engage in a process of social or behavioral comparison to estimate a relative frequency (Cole & Korkmaz, 2013). That is, when considering whether a person engages in a behavior "often," the respondent implicitly decides what or whom to compare "often" against (i.e., "often" compared to what or whom?). Potential reference groups could include, for instance, (a) other people in general, (b) same-aged peers, (c) one's friend group or social peer group, or even (d) the respondent's expectation for how often the person should (or should not) engage in the behavior. Each reference group could yield different ratings for the same absolute frequency.

In a standard questionnaire, the reference group is typically not defined, and the respondent is left to select the reference group (Wänke, 2002). Burns et al. (2001) provided an example of why this is problematic for assessing externalizing behavior: "Consider the situation where two parents indicate that the [oppositional defiant disorder] symptom argues with adults occurs very often. For one parent, very often means many times per day, while for the other parent very often may mean a few times per week. … The rating of very often probably means that the symptom occurs often enough to be a problem for both parents irrespective of the symptom's rate of occurrence … thus confounding different properties of the symptom" (p. 26).[1] In addition, people may use their own behavior as the reference of comparison. If a person engages in a behavior frequently, their standard for what is considered "often" tends to be higher, and ironically, they tend to use *lower* frequency vague quantifier terms to describe their frequency of engaging in that behavior (Goocher, 1965). This is a key weakness because it obscures individual differences and works against the goal of assessing relative frequency in a meaningful way.

In sum, interpretations of vaguely quantified reports of frequency differ based on item characteristics such as the base rate of a behavior (Pepper & Prytulak, 1974) and the composition (Chase, 1969) and order (Chan, 1991) of the other quantifiers that appear alongside the given quantifier (e.g., "often") in the same response scale (e.g., "never," "rarely," "sometimes," "very often") and respondent characteristics such as their demographic characteristics (Borgers et al., 2003; Schaeffer, 1991; Wright et al., 1994), their reference group (Wänke, 2002), their frequency of engaging in the behavior (Goocher, 1965), and their attitude toward the behavior (Goocher, 1965). Collectively, these factors may explain why conclusions based on absolute and relative frequencies sometimes differ (Schaeffer, 1991; Wänke, 2002).

Numeric Quantifiers to Assess Absolute Frequency

Given the many limitations with reports of relative frequency and vague quantifiers, it is worth considering alternatives, including reports of absolute frequency. To assess absolute frequency, researchers use numeric quantifiers. A numeric quantifier of absolute frequency could be a single, total frequency of a behavior for a given reference period (e.g., the number of times in the past month), or it could be a rate per unit of time (e.g., the number of times per day). The numeric quantifier could be specified using an open-ended response prompt (i.e., fill in the blank; Cole & Korkmaz, 2013) or using a closed-ended response prompt with a scale of category ranges (e.g., not in the past 6 months; one time in the past 6 months; two times in the past 6 months; once per month in the past 6 months; once per week in the past 6 months; once per day in the past 6 months; more than once per day in the past 6 months; Burns et al., 2001).

Weaknesses of Closed-Response Numeric Quantifiers

There are weaknesses of assessing numeric quantifiers of absolute frequency in closed-ended (Likert-type) form as opposed to in open-ended (fill-in-the-blank) form, as summarized in Supplemental Table S2. One weakness of closed-ended numeric quantifiers is that they may suppress meaningful variability, because they constrain the number of response options to the number of points on the scale (Schaeffer & Charng, 1991). A second weakness is that different items—assessing behaviors of varying frequencies—may require different frequency response options. For instance, different frequency options may be needed for detecting individual differences in frequency of arguing (higher base rate) versus physical aggression (lower base rate).

A key third limitation of closed-ended ratings is that participants indicate different frequencies based on the range of response options provided by the researchers. Low-frequency response options tend to yield lower frequency responses, whereas high-frequency response options tend to yield higher frequency responses (Schwarz et al., 1985, 1988; Tourangeau & Smith, 1996). Part of the reason for this bias may be that response options suggest normative frequencies of a behavior in a population, decreasing disclosure of higher frequency engagement in stigmatized behaviors. There is a greater impact of the particular response options (a) the more poorly the behavior is represented in memory (leading to estimation) and (b) when the behavior is ill-defined because the response options influence how readers interpret the question and their estimation strategy (Schwarz, 1999).

Measures That Assess Absolute Frequency of Externalizing Behavior

There are several examples of measures that use open-response numeric quantifiers to assess absolute frequency of externalizing behavior. Some measures use numeric quantifiers to assess behavior frequency prospectively (e.g., Student Behavior Teacher Response; Pelham et al., 2008). Tracking and counting behavior instances prospectively is commonly used in clinical practice including treatments such as parent management training (i.e., home-based tallies; Fleischman et al., 1983) and school-based interventions (e.g., tallies on a daily report card; Holdaway et al., 2020). For instance, daily report cards are used as part of an intervention in which children are rewarded for meeting behavioral goals.

There are also retrospective assessments of externalizing behavior that leverage numeric quantifiers (see Supplemental Table S3). To our knowledge, the only retrospective measures of externalizing behavior that assess absolute frequency using open-ended response

[1] Moreover, if a respondent uses one's same-aged peers as their reference group, this could have the effect of resulting in age-normed scores, which works against the goal of detecting group-level change/growth.

prompts are those that assess the number of instances that a person has used substances or engaged in delinquent behavior, as assessed by approaches such as the Timeline Follow-Back. The Timeline Follow-Back asks participants to indicate the days on which they used a substance and the quantity (or units) of the substance they used each day (Sobell & Sobell, 1992). It does not include vague quantifiers or constraints on response options as do Likert response options such as "sometimes" or "1–2 times per week." A few self-reported delinquency measures assess absolute frequency using an open-response frequency count (e.g., Bendixen et al., 2003; Elliott & Ageton, 1980). Other measures incorporate a singular item that assesses absolute frequency (e.g., Rage Attacks Questionnaire; Budman et al., 2003). However, to our knowledge, no retrospective measures of general externalizing behavior have assessed absolute frequency using open-ended response prompts.

Weaknesses of Open-Response Numeric Quantifiers

There are several weaknesses of open-ended numeric quantifiers. One weakness of numeric quantifiers is that they are relatively more difficult and burdensome than vague quantifiers due to the involvement of mathematical calculations. The greater difficulty may lead, in some cases, to longer completion time and greater item-level missingness (Lenzner et al., 2010; Schaeffer, 1991). Such difficulty may be particularly salient to higher frequency behaviors, where error may be larger. However, the response burden of numeric quantifiers can be reduced by allowing participants to select the time frame on which it is easiest to estimate the frequency of a behavior (e.g., per day, per week, per month).

Another challenge with numeric quantifiers is the complexity of scoring them. With scores from vague quantifiers, it is common to generate an aggregate scale-level score by averaging or summing scores across items. However, an average or sum score of numeric quantifiers is potentially more problematic because some behaviors are more common (and potentially less problematic) than other behaviors and thereby exert undue influence on the aggregate score. Thus, scale-level scoring of numeric quantifiers may require more complex scoring approaches, such as latent variable modeling that require larger samples.

Strengths of Open-Response Numeric Quantifiers

For all of the reasons described earlier, researchers recommend asking about absolute frequencies using an open-response format, rather than using a closed-response format (Elliott & Ageton, 1980; Schaeffer & Charng, 1991; Schaeffer & Dykema, 2020; Schaeffer & Presser, 2003; Schwarz, 1999). Numeric quantifiers address many limitations with vague quantifiers. First, numeric quantifiers have more possible response options and therefore allow greater variability in scores, allowing for stronger associations with external criteria (Cole & Korkmaz, 2012). Second, count data provide more information—that is, a greater reduction of uncertainty in the estimate of a person's level on the latent construct—than binary or polytomous items (Doebler et al., 2014).

Third, numeric quantifiers are a real metric, which leads to several strengths. The count of behavior frequency is potentially verifiable (Schaeffer, 1991)—they can be evaluated against observation or prospective tracking. As such, they are less vague, subjective, and ambiguous (Burns et al., 2001; Schaeffer, 1991). Responses to numeric quantifiers still involve some subjectivity, for example, defining what counts as a behavior occurrence. However, they are less subject to different interpretations or to confounding frequency with impairment than are vague quantifiers (Burns et al., 2001; Marincic, 2012). Indeed, numeric quantifiers have been found to reduce systematic measurement error related to different interpretations of vague quantifiers (Bradburn & Miles, 1979; Tourangeau & Smith, 1996).

Also stemming from their properties as a real metric, estimates of central tendency and variability are more meaningful. For instance, we could determine that, on average, people in their 30s drink alcohol ~2 times per month ($Mdn = 2$, $M = 6.37$, $SD = 10.26$), that 29% did not drink in the prior year, that 8% drank in the prior year but not in the past month, and that individuals at the 80th percentile drank three times per week.[2] As a real metric, numeric quantifiers do not involve subjective judgments relative to a reference group, allowing for a comparable metric across social groups (Schaeffer, 1991; Schaeffer & Dykema, 2020; Wänke, 2002).

As an additional strength of being a real metric, numeric quantifiers provide a meaningful way to assess, in the same (i.e., isomorphic) way, behaviors that are considered clinically significant at different frequencies (Burns et al., 2001). For instance, intentionally setting fires with intent to cause property damage is considered clinically significant even at a low frequency of occurrence, whereas being mean to others is considered clinically significant at a medium frequency of occurrence, and fidgeting or squirming in one's seat is considered clinically significant at a high frequency of occurrence. Meaningfully assessing—in an isomorphic way—the frequency that a person engages in each of these behaviors is not possible when using vague quantifiers because "often" means something different for each behavior (Burns et al., 2001). Absolute frequency scores are thus more meaningful, interpretable, and clinically useful than relative frequency scores.

In sum, there are important benefits to assessing absolute rather than relative frequency. Questionnaires of externalizing behavior rely largely on vague quantifiers; those that use numeric quantifiers tend to be closed-ended, both of which have limitations that prevent accurate assessment of externalizing behavior frequency. Despite longstanding recommendations to assess behavior frequency using numeric quantifiers of absolute frequency with an open-response format (Schwarz, 1999), this practice has not been widely adopted in assessment of externalizing behavior. To our knowledge, no prior study has examined an open-response frequency count of general externalizing behavior (i.e., enumeration of absolute frequency).

The Present Study

The present study examines whether assessing absolute frequency of externalizing behavior using an open-response frequency count has utility for assessment of the externalizing spectrum. We leverage modern item response theory (IRT) methods that allow estimating people's level on a latent externalizing spectrum factor using absolute frequency count data.

To evaluate the utility of the count items and modeling approach, we adapt a subset of items from the Externalizing Spectrum Inventory to use open-ended numeric quantifiers to assess absolute

[2] These are the rates in our sample for 30- to 39-year-olds in the present study.

frequency. We evaluate characteristics of the items and of the scale and validity of the latent estimates of externalizing problems. We evaluate validity with respect to interpretations of the items' scores for the intended purpose of assessing the full range of individual differences in externalizing problems, including clinical and subclinical levels. We had several hypotheses: (a) Latent externalizing scores from items assessing absolute frequency will show greater measurement precision than the Externalizing scale of the Adult Self-Report (ASR), a widely used measure of externalizing problems; latent externalizing scores will show (b) convergent validity with the ASR Externalizing scale; (c) discriminant validity with respect to internalizing problems (i.e., ASR Internalizing scale); (d) criterion validity in predicting functional impairment and inhibitory control; and (e) incremental validity in predicting functional impairment and inhibitory control above and beyond variance explained by the ASR Externalizing scale. Deficits in inhibitory control—the ability to inhibit prepotent responses—are considered an underlying phenotype of externalizing behavior (Young et al., 2009). Hypotheses were preregistered at https://osf.io/cthju (Petersen & Demko, 2024). We also test nonpreregistered secondary hypotheses: Absolute frequency items will show (f) greater measurement precision and (g) incremental validity than dichotomized versions of the same items. We also had exploratory questions, including how externalizing problems differ across ages and by sex.

Method

Participants

Participants were recruited in Fall 2023 and Spring 2024. Participants included undergraduate students in the Elementary Psychology course at the University of Iowa and people who were recruited from online sources. Students at the University of Iowa were recruited using the Sona Systems platform, and they received credits toward completion of their research exposure requirement. Volunteers were recruited from various websites, including ResearchMatch.org, the Reddit forum "r/SampleSize" (https://www.reddit.com/r/SampleSize), the Facebook group "Research/Survey" (https://www.facebook.com/groups/441242353271384/), and Twitter. The final sample included only participants recruited through Sona Systems or ResearchMatch.org. Inclusion and exclusion criteria are in Supplemental Appendix S1. A participant flowchart is in Supplemental Figure S1. To ensure legitimate responses, we included several validity checks and attention checks. The final sample included respondents who passed the validity checks and attention checks (N = 1,237; 290 males; 947 females). Participants ranged in age from 18 to 92 years of age (M = 38.36, SD = 21.18, Mdn = 30). The racial/ethnic composition of the sample is provided in Supplemental Table S4. Further description of the sample composition is in Supplemental Appendix S1.

Of the final sample, 506 (40.9%) were recruited from the University of Iowa (137 males; 369 females; M_{age} = 18.67, SD_{age} = 1.12), and 731 (59.1%) were recruited from ResearchMatch.org (153 males; 578 females; M_{age} = 51.99, SD_{age} = 17.43). The racial/ethnic composition of the University of Iowa subsample was Hispanic or Latino (12%), White (92%), Black (3%), American Indian or Alaska Native (1%), Asian (7%), Native Hawaiian or Other Pacific Islander (<1%), multiracial (6%), and other race (2%). The racial/ethnic composition of the ResearchMatch subsample was Hispanic or Latino (3%), White (92%), Black (4%), American Indian or Alaska Native (1%), Asian (4%), Native Hawaiian or Other Pacific Islander (0%), multiracial (4%), and other race (3%). Tests of systematic differences between the screened sample and the final sample are in Supplemental Appendix S2.

Procedure

Respondents completed all questionnaires online via the Research Electronic Data Capture (REDCap) platform (Harris et al., 2009). After completing the questionnaires, respondents also completed cognitive tasks. For concision, the present study focuses on the inhibitory control task given its relevance for externalizing problems.

Measures

The REDCap instrument, go/no-go task, data, a data dictionary of study variables, analysis scripts, and a computational notebook are published at https://osf.io/49m8q (Petersen et al., 2025a). Descriptive statistics and correlations of study variables are in Table 1. Tests of systematic missingness are in Supplemental Appendix S3. The study was approved by the University of Iowa Institutional Review Board (Study No.: 202305374).

Assessing the Absolute Frequency of Externalizing Behavior

Items to assess absolute frequency of externalizing behavior were largely adapted from items on the Externalizing Spectrum Inventory (Krueger et al., 2007). Not all items on the Externalizing Spectrum Inventory represent discrete, countable behaviors; for items representing behavior that is not countable, we either modified the item to be countable or dropped the item. For example, we modified "Sometimes I threaten people" to "I threaten others." We supplemented with additional items as necessary to ensure broad coverage of the externalizing spectrum. Respondents rated the frequency with which they engaged in the behavior using a 30-day reference period. They indicated, within the prior 30 days, the number of times they engaged in the behavior and the unit of time: "per day," "per week," or "in the past month." If respondents did not engage in the behavior in the past month, they indicated how many times they engaged in the behavior in the prior year. If they did not engage in the behavior in the prior year, they indicated "not in the past year." Questionnaire instructions are in Supplemental Figure S4; an example item is depicted in Supplemental Figure S5. For some low-base rate items, participants indicated how many times they engaged in the behavior in their lifetime. For examples of studies that have used a similar response format to assess the rate of various behaviors per the unit of time specified by the respondent, see Cole and Korkmaz (2013) and Peterson et al. (1982). In addition to absolute frequency items, participants also rated various other types of items, including the severity of their behavior for each item, some positive opposite items, and some items with differing response scales (e.g., Likert-type frequency, amount of credit card debt). However, the present study focused on the (non-reverse-scored) absolute frequency items so the items could be meaningfully modeled altogether.

Because frequencies were rated on different units of time, we first put all rated behaviors onto the same time scale by computing the person's rate of engaging in each behavior per year. We put items

Table 1
Correlation Matrix and Descriptive Statistics of Model Variables

Variable	1	2	3	4	5	6	7	8
1. Age	—							
2. Sex	.04	—						
3. θ	−.46***	.00	—					
4. θ (binary)	−.53***	−.01	.77***	—				
5. ASR externalizing	−.27***	.03	.56***[a]	.56***[a]	—			
6. ASR internalizing	−.23***	−.08*	.47***[b]	.41***[c]	.67***	—		
7. BFIS	−.08*	−.10***	.35***[d]	.28***[e]	.51***	.61***	—	
8. GNG	.22***	−.01	−.18***[f]	−.14***[g]	−.13***	−.07[†]	−.10*	—
N	1,237	1,237	1,237	1,237	1,180	1,180	1,181	764
Missingness (%)	0.00	0.00	0.00	0.00	4.61	4.61	4.53	38.24
M	38.36	0.23	−2.11	0.00	7.86	18.12	1.73	0.84
SD	21.18	0.42	1.11	0.98	7.28	12.61	1.64	0.10
Minimum	18.00	0.00	−7.82	−3.40	0.00	0.00	0.00	0.39
Maximum	92.00	1.00	0.08	3.19	52.00	64.00	8.75	1.00
Skewness	0.54	1.25	−1.13	−0.12	1.52	0.79	1.35	−1.03
Kurtosis	−1.22	−0.43	2.45	0.04	3.29	0.17	1.56	1.04

Note. Sex is coded as female = 0, male = 1. "θ" represents the latent externalizing problem scores from the Bayesian item response theory model of count data. "θ (binary)" represents the latent externalizing problem scores from an item response theory model based on dichotomized items. ASR = Adult Self-Report; BFIS = Barkley Functional Impairment Scale; GNG = go/no-go task.
[a] Correlations do not significantly differ from each other at $p < .05$. [b, c, d, e] Correlations significantly differ from each other at $p < .05$. [f, g] Correlations differ from each other at a trend level ($p < .10$).
* $p < .05$. *** $p < .001$. † $p < .10$.

onto the same timescale using the computeItemFrequencies() function of the petersenlab package Version 1.1 (Petersen, 2025) in R Version 4.3.1 (R Core Team, 2023). The function multiplied counts of behaviors rated "per day" by 365.25 (the number of days in a year), behaviors rated "per week" were multiplied by 365.25/7 (the number of weeks in a year), and behaviors rated "in the past month" were multiplied by 365.25/30 (the number of months in a year, standardized to 30 days per month). Items were then rounded up to the nearest integer, for the purposes of fitting the count data to a zero-inflated negative binomial model. Items rated on the lifetime scale were already in the integer form necessary for count data and were not rescaled. The final item selection process is described in Supplemental Appendix S4. The final externalizing problem items are in Supplemental Table S5. We discuss outlier handling in Supplemental Appendix S5. Internal consistency estimates (based on log-transformed counts) were $\alpha = .96$ and $\omega_{hierarchical} = .95$.

External Criteria

Behavior Problems. As an external criterion, we assessed the ASR (Achenbach & Rescorla, 2003). To evaluate convergent validity, we examined the Externalizing scale (35 items). To evaluate discriminant validity, we examined the Internalizing scale (38 items). For safety and ethical reasons, we did not assess one internalizing problem item ("I think about killing myself") because of the online administration of the questionnaires. Using the reference period of the past 6 months, participants rated items as "not true" (0), "somewhat or sometimes true" (1), or "very true or often true" (2). Internal consistency estimates were $\alpha = .90$ and $\omega_{hierarchical} = .90$ for externalizing problems and were $\alpha = .94$; $\omega_{hierarchical} = .93$ for internalizing problems. Scores were averaged across items and then multiplied by the number of items. Higher scores reflected more behavior problems. Mean *T*-scores were 48.67 (*SD* = 10.67) and 56.42 (*SD* = 12.53) for externalizing and internalizing problems, respectively.

Functional Impairment. To assess functional impairment, we used the Barkley Functional Impairment Scale (Barkley, 2011). Items assessed the degree to which people experienced impairment in various life domains during the past 6 months, ranging from "not at all" (0) to "severe" (9). We added one item to assess functional impairment relating to legal difficulties—that is, impairment/difficulties "in your history with the law/authorities (fines, arrests, jail/prison time, etc.)." Internal consistency estimates were $\alpha = .94$; $\omega_{hierarchical} = .96$. Scores were averaged across items. Higher scores reflected greater functional impairment.

Inhibitory Control. To assess inhibitory control, we used a go/no-go task. The task was administered online using jsPsych Version 7.3.3 (de Leeuw et al., 2023). The task was adapted from the Experiment Factory (Sochat et al., 2016). On a given trial, either an orange square or a blue square appeared. Participants were instructed to respond to the go stimulus—a square of a particular color—by pressing the space bar. Participants were instructed not to respond to the no-go stimulus—a square of a different color. The go and no-go stimuli were counterbalanced across participants. On a given trial, participants were given 750 ms to respond, followed by a 250 ms pause before the next trial. There were 10 practice trials. In test trials, there were 250 go trials (83.3%) and 50 no-go trials (16.7%), for a total of 300 test trials.

We took several steps to ensure measurement quality. First, the task set the participant's screen to full-screen mode; we excluded trials in which the screen was no longer in full-screen mode. Second, we excluded trials in which the reaction time was less than 200 ms, because such a response would be too fast for a person to respond deliberately to the target stimulus. Third, we dropped scores from seven participants whose percent correct on go trials was less than 60% and whose percent correct on no-go trials was less than 25%, consistent with recommendations (Congdon et al., 2012). Because

participants could receive a high score on no-go trials by performing no action, we examined the degree to which respondents inhibited a response on no-go trials and activated a response on go trials. Consistent with Eisenberg et al. (2013), the participants' inhibition score was computed by multiplying mean proportion scores from inhibition (no-go) and activation (go) trials; the final score ranged from 0 to 1. Therefore, respondents who activated a behavior on go trials but inhibited on no-go trials received the highest scores, whereas respondents who infrequently activated (or always activated) a behavior received low scores. The average split-half reliability of scores on no-go trials of 15,000 split-halves was .79. Higher scores reflected greater inhibitory control.

Statistical Analysis

First, we examined descriptive statistics of the externalizing problem items. We also examined interitem and item–total correlations, to inform item selection. In addition, we examined whether frequency ratings for each item (i.e., behavior) differed by age. We identified the best fitting form of age-related differences. Given the wide age range (18–92 years of age), we used generalized additive models (GAMs) to allow curvilinearity in the age-related estimates. To prevent overfitting, the GAMs used a *P*-spline-based smooth (i.e., penalized *B*-splines) using the gam() function of the mgcv package Version 1.9-1 (Wood, 2017) in R.

Second, we examined whether measures' scores were able to be modeled with item response modeling by examining their scores in exploratory factor analysis. The method and results of the exploratory factor analysis are in Supplemental Appendix S6.

Third, to obtain estimates of person and item parameters, we estimated Bayesian item response models, as described in Supplemental Appendix S7. To allow for nonlinearity given the wide age range (18–92 years of age), we used a GAM. Fourth, we compared measurement precision of scores on the absolute frequency items to scores on the dichotomized versions of the same items and to scores on the ASR Externalizing items. We operationalized measurement precision using item and test information. We describe the derivation of item and test information in Supplemental Appendix S8. Fifth, we examined the latent scores from the Bayesian item response model in relation to ratings of externalizing and internalizing problems on the Adult Self-Report, to evaluate convergent and discriminant validity. As a test of discriminant validity, we examined whether the latent scores were more strongly associated with ASR Externalizing than with ASR Internalizing using Fisher's *r*-to-*z* test. Sixth, we evaluated whether the latent scores showed criterion and incremental validity in predicting functional impairment and inhibitory control above and beyond (a) ASR Externalizing and (b) latent scores from IRT models fit to dichotomized items. IRT models fit to dichotomized items were estimated using the mirt package Version 1.42 (Chalmers, 2012) in R.

Results

Descriptive Statistics

Descriptive statistics of the externalizing problem items are in Supplemental Table S6. Percentiles for behavior frequencies per year of each externalizing problem item are in Supplemental Table S7. A histogram of interitem correlations of the externalizing problem items is in Supplemental Figure S6. A histogram of item–total correlations of the externalizing problem items is in Supplemental Figure S7.

We observed differing age-related courses of different behaviors. Some behaviors showed decreases in frequency across ages; other behaviors showed increases in frequency across ages, and other behaviors showed inverted-U-shaped courses in which they peaked in middle adulthood. Frequencies by age of three example behaviors are depicted in Figure 1.

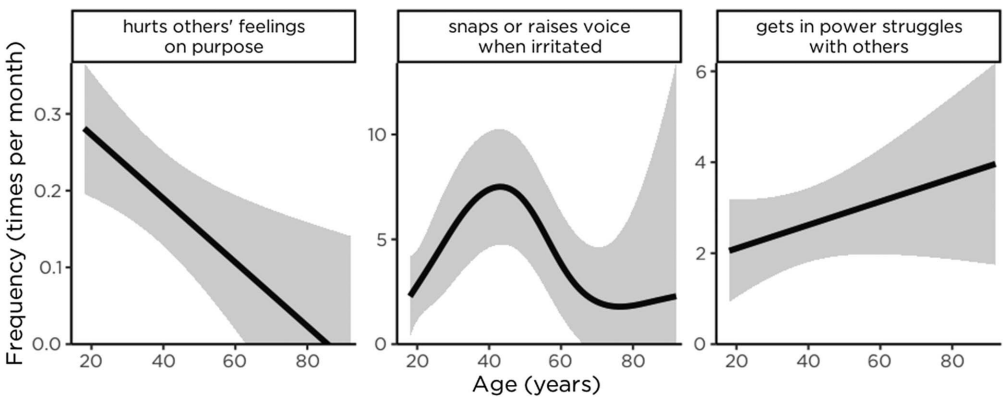
Bayesian Item Response Model

Model fit estimates from the Bayesian IRT model are in Supplemental Table S8. Regression coefficients are in Table 2. Item characteristics from the model are in Supplemental Table S9 and Figures S9–S10; item response functions are in Figure 2. Item–construct correlations are in Supplemental Table S10 and Figure S12. Model-implied estimates of externalizing problems by age and sex, from the Bayesian IRT model, are in Figure 3.

In general, externalizing problems showed strong age-related decreases. The age-related decreases were significant for women

Figure 1

Frequency (per Month) of Engaging in Various Externalizing Behaviors by Age



Note. The solid line is the model-implied estimate from a generalized additive model. The gray band is the 95% confidence interval.

Table 2
Regression Coefficients of Bayesian Item Response Theory Model

Parameter	Estimate	*SE*	Lower	Upper
Population parameters				
φ (intercept)	−2.05	0.14	−2.33	−1.80
zi (intercept)	−1.83	0.29	−2.39	−1.25
θ × Female	−2.13	0.25	−2.63	−1.64
θ × Male	−2.08	0.26	−2.58	−1.59
β (intercept)	4.24	0.42	3.45	5.09
α (intercept)	1.56	0.09	1.38	1.75
θ × Age (Shifted) × Female	−0.26	0.10	−0.43	−0.04
θ × Age (Shifted) × Male	−0.21	0.17	−0.43	0.23
Multilevel hyperparameters				
SD of φ intercept	1.25	0.12	1.04	1.51
SD of zi intercept	2.30	0.31	1.65	2.91
SD of β intercept	2.23	0.18	1.92	2.61
SD of α intercept	0.92	0.08	0.77	1.08
Smoothing spline hyperparameters				
SD of θ × Age (Shifted) × Female	0.15	0.12	0.02	0.45
SD of θ × Age (Shifted) × Male	0.12	0.15	0.00	0.55

Note. "Lower" and "Upper" represent the bounds of the 95% credible interval. Age in years was shifted (by subtracting 18) so that the value 0 represented an age of 18, the youngest age in the study. phi (φ) is the shape parameter representing the degree of overdispersion, zi is the zero-inflation parameter capturing excess zeros, theta (θ) is the person parameter representing the person's level on the latent externalizing problems factor, beta (β) is the item easiness parameter, and alpha (α) is the item discrimination parameter. *SE* = standard error.

($B = -0.26$; 95% credible interval, CI [−0.43, −0.04]) but not for men ($B = -0.21$; 95% CI [−0.43, 0.23]). Unsurprisingly, men tended to show greater externalizing problems compared to women in early adulthood (~18–30 years of age) and later adulthood (~70 years of age and older). However, surprisingly, women tended to show greater externalizing problems than men in middle adulthood (~40–60 years of age). Nevertheless, the credible intervals for men and women were largely overlapping.

Measurement Precision

Item information functions are in Supplemental Figure S12. Scores on the absolute frequency (i.e., count) items showed greater measurement precision across all levels of the construct compared to scores on the dichotomized items and the Adult Self-Report, as depicted in Figure 4. Test standard error of measurement and test reliability functions are in Supplemental Figures S13–S14.

Convergent and Discriminant Validity

Latent externalizing problem scores from the Bayesian IRT model were moderately associated with ASR Externalizing scores ($r = .56$) and were more strongly associated with ASR Externalizing than with ASR Internalizing scores ($r = .47$; $z = 4.59$, $p < .001$).

Concurrent Criterion-Related Validity

Latent externalizing problem scores from the Bayesian IRT model were moderately positively associated with functional impairment ($r = .35$, $p < .001$) and modestly negatively associated with inhibitory control ($r = -.18$, $p < .001$).

Incremental Validity

Regression coefficients of the incremental validity analyses are in Table 3. Convergent, discriminant, criterion, and incremental validity analyses by age and sex are in Supplemental Table S11. Latent externalizing problem scores from the Bayesian IRT model explained additional variance in functional impairment and inhibitory control above and beyond ASR Externalizing. ASR Externalizing was not significantly associated with inhibitory control above and beyond the latent externalizing problem scores. Latent externalizing problem scores also explained additional variance in functional impairment and inhibitory control above and beyond factor scores from an IRT model based on dichotomized items. The factor scores from the dichotomous items were not significantly associated with functional impairment or inhibitory control above and beyond the latent externalizing problem scores from the count data.

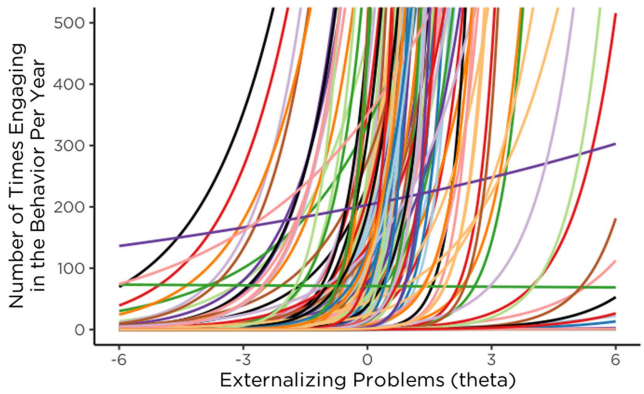
Discussion

Age-Related Courses of Externalizing Problems

In general, we observed (cross-sectional) age-related decreases in externalizing problems from early to later adulthood. However, women showed significant age-related decreases whereas men did not. Similar to prior work, we observed that men showed greater externalizing problems than women in early and older adulthood. However, interestingly, we observed that women showed relatively greater externalizing problems than men in middle adulthood (~40–60 years of age). Based on visual inspection of particular items, items that tended to show higher levels for women than men during middle adulthood included items such as snapping or raising one's voice at others when irritated, using emotional skills to make others feel guilty, and making someone feel ashamed about something they have done. One possibility is that women—on average—engage more frequently than men in such behaviors directed toward their child given their greater likelihood of assuming parental responsibilities.

We observed differing age-related courses of different behaviors, consistent with the idea that externalizing behaviors change in manifestation across development (i.e., heterotypic continuity;

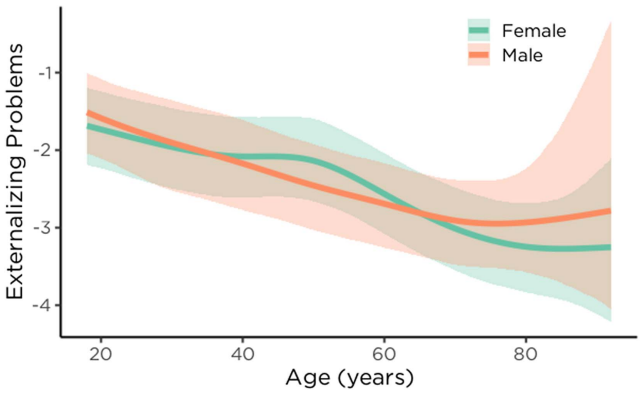
Figure 2
Item Response Functions of Items From Bayesian Item Response Model



Note. Each line represents an item. See the online article for the color version of this figure.

Figure 3

Model-Implied Latent Estimates of Externalizing Problems by Age and Sex



Note. The colored bands represent the 95% credible intervals based on the Bayesian IRT model. See the online article for the color version of this figure.

Patterson, 1993; Petersen et al., 2015). For instance, some behaviors showed decreases in frequency across ages, whereas other behaviors showed increases in frequency across ages, other behaviors showed inverted-U-shaped functions in which they peaked among middle-aged adults, and other behaviors showed no systematic differences in frequency by age. Many behaviors peaked among young adults—for example, threatening others, binge drinking, and stealing. Other behaviors peaked among middle-aged adults—for example, blaming others for one's own mistakes, avoidance or reluctance to engage in tasks that require sustained mental effort, snapping or raising voice at others, and use of cannabis and illegal drugs. Other behaviors peaked among older adults—for example, arguing and number of occasions drinking alcohol. The differing age-related courses suggest different behaviors may be especially important to assess in different developmental periods. For example, in middle adulthood, it may be especially important to assess behaviors that may be related to parenting, such as snapping or raising one's voice at others. Understanding the frequency of a wide range of externalizing behaviors in the population has critical applications for public health, for instance, knowing that parenting-related externalizing behaviors are common in middle adulthood suggests that parent training programs may usefully decrease population externalizing problems for both children and parents. In addition, the assessment of absolute frequency allowed us to quantify the number of times that people engage in a behavior per year at each of various percentiles.

Validity and Utility of Using Absolute Frequency to Assess Externalizing Problems

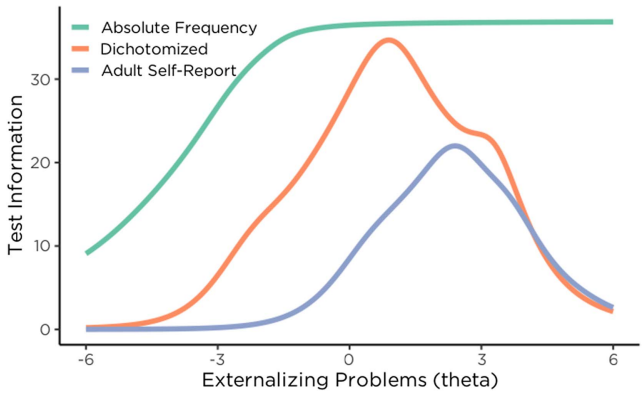
Latent scores from item response models using the items assessing absolute frequency demonstrated (a) moderate convergent validity with a widely used measure of externalizing problems (Adult Self-Report), (b) discriminant validity with respect to internalizing problems, (c) concurrent criterion validity in relation to functional impairment and inhibitory control, and (d) incremental validity in predicting functional impairment and inhibitory control above and beyond the Adult Self-Report and dichotomized versions of the same items. Moreover, scores on the items assessing absolute frequency counts yielded (e) greater information, which led to greater precision and less uncertainty in the estimates of people's level on the latent externalizing spectrum, compared to scores on the dichotomized versions of the same items and the Adult Self-Report, which assesses relative frequency. Assessment of absolute frequency showed greater precision across all levels of the construct, including at low, medium, and high/clinical levels. Thus, assessment of absolute frequency shows promise for improving assessment of the externalizing spectrum for a wide range of purposes, such as diagnosis, screening to identify at-risk individuals for purposes of secondary prevention, and—consistent with arguments by Elliott and Ageton (1980)—evaluation of the full range of individual differences.

With few exceptions, findings of convergent, discriminant, criterion, and incremental validity were supported when examining the sample disaggregated by age and sex, providing further confidence in the findings. Nevertheless, the latent scores showed somewhat weaker discriminant validity among older adults, and incremental validity (above ASR Externalizing and dichotomized versions of the same items) was somewhat weaker in men.

An additional factor that may artificially inflate correlations is similarity in response format between measures, a form of shared method bias (Podsakoff et al., 2012). Because both the ASR and the functional impairment criterion (Barkley Functional Impairment Scale) use vague quantifiers, this may lead to spuriously higher criterion correlations than measures using numeric quantifiers. Consistent with this possibility, the ASR showed stronger criterion correlations (than absolute frequency) with functional impairment (both using vague quantifiers), whereas the absolute frequency items (which used numeric quantifiers) showed stronger criterion correlations (than the ASR) with inhibitory control (which was assessed with a numeric metric). This possibility warrants further investigation.

Figure 4

Test Information Functions of Absolute Frequency Versus Dichotomized Versus Adult Self-Report Items: Test Information as a Function of Theta (θ)



Note. "Absolute Frequency" refers to count items from the Bayesian item response model; "Dichotomized" refers to dichotomized versions of the count items that were fit to an item response theory model; "Adult Self-Report" refers to items from the Adult Self-report that were fit to an item response theory model. "θ" represents the person's level on the latent externalizing problems factor. The Adult Self-Report includes fewer items than the absolute frequency/dichotomized count items. See the online article for the color version of this figure.

Table 3
Regression Coefficients From Incremental Validity Analyses

Predictor	*B*	β	*SE*	*p*
Outcome: functional impairment				
Incremental validity versus ASR				
ASR externalizing	0.102	0.45	0.007	<.001
θ (Count)	0.140	0.09	0.045	.001
Incremental validity versus θ (Binary)				
θ (Binary)	0.031	0.02	0.071	.668
θ (Count)	0.494	0.33	0.063	<.001
Outcome: inhibitory control				
Incremental validity versus ASR				
ASR externalizing	0.000	−0.02	0.001	.593
θ (Count)	−0.009	−0.10	0.004	.049
Age (Shifted)	0.008	0.17	0.002	<.001
Incremental validity versus θ (Binary)				
θ (Binary)	0.007	0.07	0.006	.258
θ (Count)	−0.014	−0.15	0.005	.007
Age (Shifted)	0.009	0.19	0.002	<.001

Note. Bold coefficients are significant at $p < .05$ level. "θ (Count)" represents the latent externalizing problem scores from the Bayesian item response theory model of count data. "θ (Binary)" represents the latent externalizing problem scores from an item response theory model based on dichotomized items. The models predicting inhibitory control include age as a covariate because of the known developmental changes in inhibitory control. Age in years was shifted (by subtracting 18) so that the value 0 represented an age of 18, the youngest age in the study. *SE* = standard error; ASR = Adult Self-Report.

Our finding of incremental validity of absolute frequency over dichotomous items extends findings from Bendixen et al. (2003) that a (log-transformed) absolute frequency scale outperformed—but only modestly—a sum of the (dichotomous) presence/absence of various behaviors (i.e., a variety scale) in prediction. However, Bendixen et al. did not use a model specifically for count data (e.g., negative binomial); instead, they used a sum score, thus giving greater weight to higher frequency—but potentially less severe—behaviors. By contrast, our model accounted for the fact that different behaviors had different base rates and weighted behaviors according to how strongly each was associated with the latent externalizing factor.

Evidence has shown that relative frequencies are poor substitutes for absolute frequencies (Schaeffer, 1991). Given the importance of knowing the absolute frequency of a behavior, it can be valuable to ask it directly. Nevertheless, just because respondents provide a more precise response (with a numeric quantifier as opposed to a vague quantifier) does not necessarily mean it is accurate; apparent precision can create a false sense of accuracy (Wright et al., 1994). There can be biases in responding including perceptual and recounting biases (Laursen et al., 2012).

We do not intend to discard vague quantifiers entirely—they can have utility. Nevertheless, absolute frequency quantifiers likely provide incremental utility for assessment of externalizing behavior. Indeed, some have noted that, because "behavioral frequency reports are error-prone anyway, why bother asking respondents for reports that suggest more precision than they can provide? Unfortunately, vague frequency expressions carry their own load of problems … different respondents use the same term to mean different objective frequencies of the same behavior" (Sudman et al., 1996, p. 226). As noted by Schaeffer (1991), "despite their problems, absolute frequencies avoid the ambiguities of relative frequencies" (p. 416). Schwarz (1999) described use of vague quantifiers to assess behavior frequency as "the worst possible choice" (p. 99).

Findings are mixed regarding whether vague or numeric quantifiers outperform the other in prediction; some studies have preferred numeric quantifiers (for a review, see Bradburn & Miles, 1979); others have preferred vague quantifiers (Al Baghal, 2014a, 2014b). The utility of one over the other may depend on the behavior assessed. Vague quantifiers may be preferred when rating the frequency of behaviors that are regular, similar, and relatively mundane or emotionally neutral over a longer reference period (Marincic, 2012). Enumerating (counting) the frequency of events (e.g., behaviors) is accurate when the event is salient/severe and infrequent and when the period of recall is shorter (Laursen et al., 2012; Marincic, 2012). When the event is more frequent or more mundane and lacks distinguishing features, forgetting and underreporting are a greater risk (Laursen et al., 2012; Marincic, 2012). Enumeration strategies (as opposed to estimation) can be enhanced by (a) asking participants to count specific events instead of estimating rates/frequencies, (b) alerting respondents about the task difficulty, and (c) reminding respondents about the importance of accurate reports (Laursen et al., 2012).

Strengths, Limitations, and Constraints on Generality

The study had limitations. First, the sample was not nationally representative. Thus, the frequency estimates for engaging in various behaviors may not accurately reflect national norms. Because all participants were from the United States, the utility of the items may differ in other countries and cultures. Second, the items were assessed via self-report, which may lead to social desirability bias and other reporting biases. Third, the ratings were made retrospectively. When comparing retrospective and prospective ratings of substance use, some studies have observed underreporting when using retrospective ratings (Poulton et al., 2018; Yang et al., 2023); one study observed overreporting (Willis et al., 2021). Fourth, the study was cross-sectional; thus, the observed age-related differences could reflect cohort or sampling differences (e.g., college students vs. online recruitment) rather than differences arising as a function of development. Fifth, the Bayesian item response model assumes unidimensionality of the externalizing spectrum. It will be valuable for future work to examine multidimensional models (e.g., Magnus & Garnier-Villarreal, 2022). Sixth, we did not include all items from the Externalizing Spectrum Inventory because not all items could be easily converted to assess absolute frequency in a countable way. Thus, the assessment may have content gaps. Seventh, the items were modeled in terms of rate per year. However, the number of times in the past 30 days that a person engaged in a behavior may not extrapolate to the number of times that they engaged in the behavior in the past year. Eighth, the present study did not directly compare assessments of absolute versus relative frequency. This will be an important direction for future research.

The study also had key strengths. First, the study assessed absolute frequency—this is the first study to assess open-response frequency counts of general externalizing behavior in which the respondent is able to specify the most appropriate time frame for each behavior. This allowed determining the frequency with which each behavior occurred. Second, the study spanned a wide age range and included older adults, unlike many studies that focus on college

students. Third, we applied advanced modeling of the items to obtain latent externalizing problem scores. To our knowledge, this is the first study to apply an IRT model specifically designed for count data to externalizing problems—a key step toward scoring individuals' levels of externalizing behavior based on absolute frequency. Fourth, we examined a performance-based task, which removes shared method biases. Fifth, we examined multiple aspects of validity, finding that scores on the assessment of absolute frequency showed convergent, discriminant, criterion, and incremental validity for assessing the externalizing spectrum. Sixth, we compared count and dichotomous versions of the same items and found that scores on the count data led to greater precision and stronger criterion-related validity compared to scores on the dichotomized items.

Conclusions and Future Directions

The assessment of absolute frequency in the present study led to important innovations. It allowed us to determine how common each behavior is. Latent scores derived from items assessing absolute frequency demonstrated convergent, discriminant, criterion, and incremental validity for assessing the externalizing spectrum, and they demonstrated improved precision compared to dichotomous items. In sum, our findings collectively suggest there is utility in assessing absolute frequency of externalizing behavior. Although the present study did not directly compare assessments of absolute versus relative frequency, our findings suggest that assessment of absolute frequency of externalizing behavior provides a more useful way to assess the externalizing spectrum (Burns et al., 2001). It will be important for future work to integrate frequency information with information about other aspects of the behavior, including its intensity, duration, and the impact of the behavior and resulting impairment. In addition, future work should compare absolute versus relative frequencies directly in terms of their utility for assessment of the externalizing spectrum. Finally, we expect that assessments of absolute frequency will also be valuable for other constructs, in addition to externalizing behavior.

References

Achenbach, T. M., & Rescorla, L. A. (2003). *Manual for the ASEBA adult forms & profiles*. University of Vermont, Research Center for Children, Youth, & Families.

Al Baghal, T. (2014a). Is vague valid? The comparative predictive validity of vague quantifiers and numeric response options. *Survey Research Methods*, *8*(3), 169–179. https://doi.org/10.18148/srm/2014.v8i3.5813

Al Baghal, T. (2014b). Numeric estimation and response options: An examination of the accuracy of numeric and vague quantifier responses. *Journal of Methods and Measurement in the Social Sciences*, *5*(2), Article 18. https://doi.org/10.2458/v5i2.18476

Barkley, R. A. (2011). *Barkley functional impairment scale (BFIS)*. Guilford Press. https://www.guilford.com/books/Barkley-Functional-Impairment-Scale-BFIS-for-Adults/Russell-Barkley/9781609182199

Batchelor, J. H., & Miao, C. (2016). Extreme response style: A meta-analysis. *Journal of Organizational Psychology*, *16*(2), 51–62. https://www.researchgate.net/publication/316820164_Extreme_Response_Style_A_Meta-Analysis (archived at https://perma.cc/L54G-ZMMG)

Bendixen, M., Endresen, I. M., & Olweus, D. (2003). Variety and frequency scales of antisocial involvement: Which one is better? *Legal and Criminological Psychology*, *8*(2), 135–150. https://doi.org/10.1348/135532503322362924

Borgers, N., Hox, J., & Sikkel, D. (2003). Response quality in survey research with children and adolescents: The effect of labeled response options and vague quantifiers. *International Journal of Public Opinion Research*, *15*(1), 83–94. https://doi.org/10.1093/ijpor/15.1.83

Bradburn, N. M., & Miles, C. (1979). Vague quantifiers. *Public Opinion Quarterly*, *43*(1), 92–101. https://doi.org/10.1086/268494

Budman, C. L., Rockmore, L., Stokes, J., & Sossin, M. (2003). Clinical phenomenology of episodic rage in children with Tourette syndrome. *Journal of Psychosomatic Research*, *55*(1), 59–65. https://doi.org/10.1016/S0022-3999(02)00584-6

Burns, G. L., Walsh, J. A., Patterson, D. R., Holte, C. S., Sommers-Flanagan, R., & Parker, C. M. (2001). Attention deficit and disruptive behavior disorder symptoms: Usefulness of a frequency count rating procedure to measure these symptoms. *European Journal of Psychological Assessment*, *17*(1), 25–35. https://doi.org/10.1027//1015-5759.17.1.25

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chan, J. C. (1991). Response-order effects in Likert-type scales. *Educational and Psychological Measurement*, *51*(3), 531–540. https://doi.org/10.1177/0013164491513002

Chase, C. I. (1969). Often is where you find it. *American Psychologist*, *24*(11), Article 1043. https://doi.org/10.1037/h0037829

Cole, J. S., & Korkmaz, A. (2012). *Estimation of expected academic engagement behaviors: The use of vague quantifiers versus tallied responses*. Paper presented at the annual meeting of the American Educational Research Association, Vancouver, British Columbia, Canada. https://hdl.handle.net/2022/23831

Cole, J. S., & Korkmaz, A. (2013). Estimating college student behavior frequencies: Do vague and enumerated estimation strategies yield similar results? *Journal of Applied Research in Higher Education*, *5*(1), 58–71. https://doi.org/10.1108/17581181311310270

Congdon, E., Mumford, J. A., Cohen, J. R., Galvan, A., Canli, T., & Poldrack, R. A. (2012). Measurement and reliability of response inhibition. *Frontiers in Psychology*, *3*, Article 37. https://doi.org/10.3389/fpsyg.2012.00037

de Leeuw, J. R., Gilbert, R. A., & Luchterhandt, B. (2023). jsPsych: Enabling an open-source collaborative ecosystem of behavioral experiments. *Journal of Open Source Software*, *8*(85), Article 5351. https://doi.org/10.21105/joss.05351

Doebler, A., Doebler, P., & Holling, H. (2014). A latent ability model for count data and application to processing speed. *Applied Psychological Measurement*, *38*(8), 587–598. https://doi.org/10.1177/0146621614543513

Eisenberg, N., Edwards, A., Spinrad, T. L., Sallquist, J., Eggum, N. D., & Reiser, M. (2013). Are effortful and reactive control unique constructs in young children? *Developmental Psychology*, *49*(11), 2082–2094. https://doi.org/10.1037/a0031745

Elliott, D. S., & Ageton, S. S. (1980). Reconciling race and class differences in self-reported and official estimates of delinquency. *American Sociological Review*, *45*(1), 95–110. https://doi.org/10.2307/2095245

Fleischman, M. J., Horne, A. M., & Arthur, J. L. (1983). *Troubled families: A treatment program*. Research PressPub.

Goocher, B. E. (1965). Effects of attitude and experience on the selection of frequency adverbs. *Journal of Verbal Learning and Verbal Behavior*, *4*(3), 193–195. https://doi.org/10.1016/S0022-5371(65)80020-2

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics*, *42*(2), 377–381. https://doi.org/10.1016/j.jbi.2008.08.010

Holdaway, A. S., Hustus, C. L., Owens, J. S., Evans, S. W., Coles, E. K., Egan, T. E., Himawan, L., Zoromski, A. K., Dawson, A. E., & Mixon, C. S. (2020). Incremental benefits of a daily report card over time for youth with disruptive behavior: Replication and extension. *School Mental Health*, *12*(3), 507–522. https://doi.org/10.1007/s12310-020-09375-w

Krueger, R. F., Markon, K. E., Patrick, C. J., Benning, S. D., & Kramer, M. D. (2007). Linking antisocial behavior, substance use, and personality: An integrative quantitative model of the adult externalizing spectrum. *Journal of Abnormal Psychology*, *116*(4), 645–666. https://doi.org/10.1037/0021-843X.116.4.645

Laursen, B. P., Denissen, J., & Bjorklund, D. F. (2012). Event frequency measurement. In B. P. Laursen, T. D. Little, & N. A. Card (Eds.), *Handbook of developmental research methods* (pp. 66–81). Guilford Press.

Lenzner, T., Kaczmirek, L., & Lenzner, A. (2010). Cognitive burden of survey questions and response times: A psycholinguistic experiment. *Applied Cognitive Psychology*, *24*(7), 1003–1020. https://doi.org/10.1002/acp.1602

Magnus, B. E., & Garnier-Villarreal, M. (2022). A multidimensional zero-inflated graded response model for ordinal symptom data. *Psychological Methods*, *27*(2), 261–279. https://doi.org/10.1037/met0000395

Marincic, J. L. (2012). *Measurement models for behavioral frequencies: A comparison between numerically and vaguely quantified reports*. Mathematica policy research, Working paper 10. https://econpapers.repec.org/paper/mprmprres/a55a64fc06fb4fc4af6621cb9523179b.htm (archived at https://web.archive.org/web/20240916202802/https://www.mathematica.org/-/media/publications/pdfs/measurement_models_wp.pdf)

Patterson, G. R. (1993). Orderly change in a stable world: The antisocial trait as a chimera. *Journal of Consulting and Clinical Psychology*, *61*(6), 911–919. https://doi.org/10.1037/0022-006X.61.6.911

Pelham, W. E., Greiner, A. R., & Gnagy, E. M. (2008). *Student behavior teacher response observation code manual* [Unpublished manual]. Department of Psychology, University at Buffalo.

Pepper, S., & Prytulak, L. S. (1974). Sometimes frequently means seldom: Context effects in the interpretation of quantitative expressions. *Journal of Research in Personality*, *8*(1), 95–101. https://doi.org/10.1016/0092-6566(74)90049-X

Petersen, I. T. (2025). *petersenlab: A collection of R functions by the Petersen Lab* [R package]. Comprehensive R Archive Network. https://doi.org/10.32614/CRAN.package.petersenlab

Petersen, I. T., Bates, J. E., Dodge, K. A., Lansford, J. E., & Pettit, G. S. (2015). Describing and predicting developmental profiles of externalizing problems from childhood to adulthood. *Development and Psychopathology*, *27*(3), 791–818. https://doi.org/10.1017/S0954579414000789

Petersen, I. T., & Demko, Z. (2024). *Pre-registration of hypotheses in the developing a novel measure of the externalizing spectrum—Pilot study*. https://doi.org/10.17605/OSF.IO/GSRD2

Petersen, I. T., Demko, Z., Doebler, P., Sabel, L., Oleson, J. J., & Krueger, R. F. (2025a). *Data, analysis code, and research materials for "How often is 'often'? Improving assessment of the externalizing spectrum using absolute frequency"*. https://doi.org/10.17605/OSF.IO/49M8Q

Petersen, I. T., Demko, Z., Doebler, P., Sabel, L., Oleson, J. J., & Krueger, R. F. (2025b). *How often is "often"? Improving assessment of the externalizing spectrum using absolute frequency*. PsyArXiv. https://doi.org/10.31234/osf.io/r9vqz_v3

Peterson, M. A., Honig, P. K., Chaiken, J. M., & Ebener, P. A. (1982). *Survey of prison and jail inmates: Background and method*. RAND Corporation.

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology*, *63*(1), 539–569. https://doi.org/10.1146/annurev-psych-120710-100452

Poulton, A., Pan, J., Bruns, L. R., Jr., Sinnott, R. O., & Hester, R. (2018). Assessment of alcohol intake: Retrospective measures versus a smartphone application. *Addictive Behaviors*, *83*, 35–41. https://doi.org/10.1016/j.addbeh.2017.11.003

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org

Schaeffer, N. C. (1991). Hardly ever or constantly? Group comparisons using vague quantifiers. *Public Opinion Quarterly*, *55*(3), 395–423. https://doi.org/10.1086/269270

Schaeffer, N. C., & Charng, H.-W. (1991). Two experiments in simplifying response categories: Intensity ratings and behavioral frequencies. *Sociological Perspectives*, *34*(2), 165–182. https://doi.org/10.2307/1388989

Schaeffer, N. C., & Dykema, J. (2020). Advances in the science of asking questions. *Annual Review of Sociology*, *46*(1), 37–60. https://doi.org/10.1146/annurev-soc-121919-054544

Schaeffer, N. C., & Presser, S. (2003). The science of asking questions. *Annual Review of Sociology*, *29*(1), 65–88. https://doi.org/10.1146/annurev.soc.29.110702.110112

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, *54*(2), 93–105. https://doi.org/10.1037/0003-066X.54.2.93

Schwarz, N., Hippler, H.-J., Deutsch, B., & Strack, F. (1985). Response scales: Effects of category range on reported behavior and comparative judgments. *Public Opinion Quarterly*, *49*(3), 388–395. https://doi.org/10.1086/268936

Schwarz, N., Strack, F., Müller, G., & Chassein, B. (1988). The range of response alternatives may determine the meaning of the question: Further evidence on informative functions of response alternatives. *Social Cognition*, *6*(2), 107–117. https://doi.org/10.1521/soco.1988.6.2.107

Sobell, L. C., & Sobell, M. B. (1992). Timeline follow-back. In R. Z. Litten & J. P. Allen (Eds.), *Measuring alcohol consumption: Psychosocial and biochemical methods* (pp. 41–72). Humana Press. https://doi.org/10.1007/978-1-4612-0357-5_3

Sochat, V. V., Eisenberg, I. W., Enkavi, A. Z., Li, J., Bissett, P. G., & Poldrack, R. A. (2016). The experiment factory: Standardizing behavioral experiments. *Frontiers in Psychology*, *7*, Article 610. https://doi.org/10.3389/fpsyg.2016.00610

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. Jossey-Bass.

Tourangeau, R., & Smith, T. W. (1996). Asking sensitive questions: The impact of data collection mode, question format, and question context. *Public Opinion Quarterly*, *60*(2), 275–304. https://doi.org/10.1086/297751

Wänke, M. (2002). Conversational norms and the interpretation of vague quantifiers. *Applied Cognitive Psychology*, *16*(3), 301–307. https://doi.org/10.1002/acp.787

Willis, M., Marcantonio, T. L., & Jozkowski, K. N. (2021). Momentary versus retrospective reports of alcohol or cannabis use, sexual activity, and their co-occurrence. *Addictive Behaviors*, *119*, Article 106932. https://doi.org/10.1016/j.addbeh.2021.106932

Wirt, R. D., Lachar, D., Seat, P. D., & Broen, W. E. (2001). *Personality inventory for children* (2nd ed.). Western Psychological Services.

Wood, S. N. (2017). *Generalized additive models: An introduction with R* (2nd ed.). CRC Press. https://doi.org/10.1201/9781315370279

Wright, D. B., Gaskell, G. D., & O'Muircheartaigh, C. A. (1994). How much is 'quite a bit'? Mapping between numerical values and vague quantifiers. *Applied Cognitive Psychology*, *8*(5), 479–496. https://doi.org/10.1002/acp.2350080506

Yang, J. J., Ou, T.-S., Lin, H.-C., Kyung Nam, J., Piper, M. E., & Buu, A. (2023). Retrospective and real-time measures of the quantity of e-cigarette use: An ecological momentary assessment study. *Nicotine & Tobacco Research*, *25*(10), 1667–1675. https://doi.org/10.1093/ntr/ntad094

Young, S. E., Friedman, N. P., Miyake, A., Willcutt, E. G., Corley, R. P., Haberstick, B. C., & Hewitt, J. K. (2009). Behavioral disinhibition: Liability for externalizing spectrum disorders and its genetic and environmental relation to response inhibition across adolescence. *Journal of Abnormal Psychology*, *118*(1), 117–130. https://doi.org/10.1037/a0014657

Received February 11, 2025
Revision received October 8, 2025
Accepted October 15, 2025 ■

Supplementary Appendix S1. Description of Participants.

Inclusion criteria include: (1) the person is 18 years old or older and (2) the person lives in the U.S. Exclusion criteria include: (1) the person is unable to read or write proficiently in English, (2) the person does not have normal or corrected-to-normal vision, and (3) the person fails an attention or validity check.

In terms of gender, 903 (73.0%) identified as female, 295 (23.8%) identified as male, 33 (2.7%) identified as non-binary, and 6 (0.5%) identified as "Other" gender. A histogram of participants' ages is in Supplementary Figure S2. In terms of marital status, 720 (58.2%) were single/never married, 337 (27.2%) were married, 9 (0.7%) were separated, 120 (9.7%) were divorced, 23 (1.9%) were re-married, and 28 (2.2%) were widowed or a widower. In terms of educational attainment, 21 (1.7%) had some high school, 250 (20.2%) were high school graduates, 371 (30.0%) had some college, 62 (5.0%) had an associate degree, 234 (18.9%) had a bachelor's degree, 223 (18.0%) had a master's degree, 25 (2.0%) had a professional school degree, and 51 (4.1%) had a doctoral degree. Median family income was $99,000 (*SD* = $308,071; median absolute deviation = $75,613). A histogram of participants' family incomes is in Supplementary Figure S3. In summary, compared to the U.S. population, participants in the sample were more likely to be female, Non-Hispanic White, single/never married, middle or upper class, and have a college or graduate degree. However, participant demographics were broadly reflective of the recruitment sources.

Supplementary Appendix S2. Tests of Systematic Differences Between the Screened Sample and the Final Sample.

We examined whether there were systematic differences between the screened sample and the final sample in terms of demographic characteristics, socioeconomic status, and the recruitment source. Where systematic differences were observed, we also examined whether there were differences between the screened sample and the final sample among the ResearchMatch participants or the University of Iowa student participants, separately.

Compared to the screened sample, participants in the final sample were more likely to have been recruited from the University of Iowa Sona system, as opposed to from ResearchMatch, Twitter, Facebook, or Reddit. Although we recruited participants from a variety of online sources in addition to Sona, including ResearchMatch, Twitter, Facebook, and Reddit, no participants in the final sample were from Twitter, Facebook, or Reddit.

There were no differences between the screened and final sample in terms of the participant's sex or gender. In terms of age, participants in the final sample were more likely to be younger (M = 38.17 years) compared to the screened sample (M = 45.62 years; t[3,755.4] = 7.20, $p < .001$). This appears to be due to a higher rate of survey completion among the college student subsample than for the ResearchMatch subsample. Among the ResearchMatch subsample, the final sample tended to be older (M = 52.20 years) than the screened sample (M = 47.30 years; t[2798.6] = −4.48, $p < .001$). There were no age-related differences between the screened and final sample for the college student subsample.

In terms of the participant's race/ethnicity, compared to the screened sample (87%), participants in the final sample were more likely to be White (92%; $\chi^2[1] = 11.67$, $p < .001$), a difference which was at a trend-level in the ResearchMatch subsample ($\chi^2[1] = 3.16$, $p = .076$)

and was significant in the college student subsample ($\chi^2[1] = 20.17$, $p < .001$). Compared to the screened sample (8%), participants in the final sample were less likely to be Black (3%; $\chi^2[1] = 19.51$, $p < .001$), a difference which held in both the ResearchMatch ($\chi^2[1] = 7.89$, $p = .005$) and college student ($\chi^2[1] = 8.95$, $p = .003$) subsamples. There were no differences between the screened and final sample in terms of the participant's Hispanic/Latino ethnicity, American Indian, Asian, Arab/Middle Eastern/North African, Native Hawaiian or Other Pacific Islander, Multiracial, or Other Race.

In terms of education, compared to the screened sample, participants in the final sample tended to have lower education ($\chi^2[8] = 105.32$, $p < .001$), which could reflect the higher rate of survey completion among the college student subsample than for the ResearchMatch subsample. There were no education-related differences between the screened and final sample for the college student or ResearchMatch subsamples.

In terms of income, compared to the screened sample, participants in the final sample tended to have higher income ($t[1,597.3] = -2.29$, $p = .022$). This difference held in the ResearchMatch subsample ($t[1,305.3] = -2.56$, $p = .011$) but not in the college student subsample.

In sum, compared to the screened sample, participants in the final sample were somewhat more likely to be recruited from the University of Iowa (than from other online sources), to be older (among the ResearchMatch subsample), to be White and not Black, and to have higher education and income (among the ResearchMatch subsample). There were no differences between the screened and final sample in terms of the participant's sex, gender, Hispanic/Latino ethnicity, American Indian, Asian, Arab/Middle Eastern/North African, Native Hawaiian or Other Pacific Islander, Multiracial, or Other Race.

Supplementary Appendix S3. Tests of Systematic Missingness.

We evaluated whether there was systematic missingness in the model variables. Among those who passed the validity and attention checks, there was essentially no missingness in scores on the externalizing problem items. Thus, for tests of systematic missingness, we focus on the other assessments: Adult Self-Report, Barkley Functional Impairment System, and go/no-go task. For the Adult Self-Report and Barkley Functional Impairment System, there was no systematic missingness as a function of sex, gender, race, ethnicity, education, or family income. However, there was some systematic missingness as a function of age and recruitment source. For the Adult Self-Report and Barkley Functional Impairment System, older adults and individuals recruited from ResearchMatch were more likely to be missing scores compared to younger adults (ASR: $t[61.87] = -2.14$, $p = .036$; BFIS: $t[60.46] = -1.95$, $p = .056$) and compared to individuals recruited through SONA Systems (ASR: $\chi^2[1] = 14.52$, $p < .001$; BFIS: $\chi^2[1] = 11.91$, $p < .001$). This may be partially attributable to the lack of compensation for participants recruited online, whereas college students recruited from the University of Iowa received credits toward completion of their research exposure requirement. For the go/no-go task, there was no systematic missingness as a function of age, race, ethnicity, family income, or recruitment source. However, there was some systematic missingness as a function of sex/gender ($\chi^2[1] = 5.77$, $p = .016$) and education ($\chi^2[7] = 20.47$, $p = .005$), with females and those with lower levels of education more likely to be missing scores than males and those with higher levels of education.

We performed Little's (1988) missing completely at random (MCAR) test to determine whether our data missingness significantly deviates from a missing completely at random pattern. To perform the MCAR test, we used the mcar_test() function of the naniar package

version 1.1.0 (Tierney & Cook, 2023) in R. The MCAR test was nonsignificant ($\chi^2[22] = 25.50$, $p = .274$), indicating that the data missingness did not significantly deviate from a missing completely at random pattern.

Supplementary Appendix S4. Final Item Selection.

For final item selection, we removed six items that had little or no variability ("generate or post online sexual content involving someone without their knowledge and/or explicit consent"; "threaten to tell others a secret or lie to convince someone to have sex"; "harm or threaten to harm someone physically to convince them to have sex"; "rob a bank, store, or other business"; "embezzle money"; "slip someone drugs so that I could take advantage of them"). We also removed one item that had weak associations with the total score and with the latent factor ("How many sexual partners have you had in your life?"). Although a few remaining items showed relatively modest associations with the total score and with the latent factor, they were retained given their strong conceptual relevance for the construct of externalizing problems (e.g., "forced someone into sexual activity"; "deliberately started fires that caused damage"). The final item set included 127 items.

Supplementary Appendix S5. Outlier Handling.

Because we used an open-ended response format (rather than a closed-response format), we had some outlying responses. We winsorized outliers so that the outlying responses did not receive undue weight in the modeling. Because observations were on a nonnormally distributed count scale, we used an approach to outlier detection for skewed distributions. We used the adjboxStats() function from the robustbase package version 0.99-2 (Maechler et al., 2024) in R to identify outliers. The adjboxStats() function identifies the upper fence (upper whisker) of a box-and-whisker plot for skewed distributions based on methods proposed by Hubert and Vandervieren (2008). To account for the extent of skewness, the method uses the medcouple, which is a robust measure of the extent to which data deviate from a symmetric distribution (i.e., are skewed). Instead of the traditional boxplot (where the upper fence is defined as $Q_3 + 1.5 \cdot$ IQR), the upper fence of the adjusted boxplot is defined as:

$$Q_3 + c \cdot e^{b \cdot M} \cdot \text{IQR} \quad (1)$$

where Q_3 is the third quartile, c is a coefficient that determines how far the whiskers extend out from the box of the boxplot, b is a scaling factor that is mutliplied by the medcouple to determine the outlier boundaries, M is the medcouple, and IQR is the interquartile range ($Q_3 - Q_1$). We used the default coefficients for identification of outliers ($c = 1.5$; $b = 3$). We examined the upper fence for each item under two conditions. First, we identified the upper fence for each item without excluding any count responses. Given the large proportion of 0s and 1s in the count data, the many 0/1 responses could greatly lower the threshold for what is considered an outlier. Thus, second, we identified the upper fence for each item when excluding 0 and 1 responses (i.e., 0 or 1 times per year). Based on these two upper fences for each item, to be conservative when identifying outliers, we selected the upper fence as the highest of these two

fence values for each item (i.e., the higher upper fence of the following: the upper fence of the raw data and the upper fence when excluding 0/1 responses). We then identified outliers as values that were greater than the upper fence for that item. To winsorize the outliers, we identified the highest value below the upper fence for each item. We then truncated the outliers by setting them to the highest value below the upper fence for that item and added 1 (so the response was slightly larger than the highest non-outlier, to retain order of individual differences in frequency).

Supplementary Appendix S6. Exploratory Factor Analysis.

We first examined whether measures' scores were able to be modeled with item response modeling by examining their scores in exploratory factor analysis (EFA). We conducted EFA using the psych package version 2.4.6.26 (Revelle, 2021) in R.

We examined the externalizing problem items in EFA. Results of the EFA are in Table S10. Most items had a standardized factor loading above .20 on the latent externalizing problems factor. A one-factor model accounted for 17% of the variance. In a two-factor model, the second factor accounted for 4% of the variance. The ratio of the first eigenvalue to the second eigenvalue was greater than 4 (4.06), suggesting that the data are unidimensional enough for estimating a common latent factor (Slocum-Gori & Zumbo, 2011). Thus, although the externalizing problem items clearly assessed multiple dimensions, a single factor accounted for considerable variance, and accounted for considerably more variance than the second factor. Based on this evidence, the primary factor appeared to reflect a meaningful operationalization of externalizing problems. Thus, given our goals to examine a latent factor of externalizing problems, we conducted item response modeling with a single factor.

Supplementary Appendix S7. Bayesian Item Response Theory Model.

We used a log-linear count data model to build an explanatory item response theory (IRT) model (i.e., an IRT model that incorporates predictors). We estimated the IRT model in a Bayesian mixed modeling framework using the brms package version 2.21.0 (Bürkner, 2017) in R. The unidimensional nonlinear IRT model allows us to simultaneously generate estimates of item and scale utility and to model the age-related course of externalizing problems. In the present study, externalizing problems ratings were count data. Due to overdispersion, with few exceptions, the conditional variance of items' scores was greater than the mean. Thus, we used a negative binomial response distribution instead of a Poisson distribution (Hu et al., 2011). Because many scores were zero, we used a zero-inflated negative binomial distribution for the outcome variable. A zero-inflated model assumes that the ratings of zero result from a mixture of two distributions (i.e., have two different origins): structural or sampling (Hu et al., 2011). The sampling zeros reflect the usual negative binomial distribution, which assumes that the ratings of zero happened by chance. By contrast, the structural zeros reflect a particular structure of the data that leads to excessive zeros above and beyond zeros due to random chance. Participants who are at risk of externalizing behavior but who report not having engaged in externalizing behavior during the given timeframe are considered sampling zeros. Participants who are not at risk of externalizing behavior—e.g., those who do not have the opportunity to engage in a given behavior—are considered structural zeros.

To account for the excess zeros, the IRT model estimates four parameters for each item, item j. Similar to a binary IRT model with intercept-slope parametrization (Beisemann, 2022), we include an easiness parameter (β_j; i.e., beta, similar to an intercept) and a discrimination parameter (α_j; i.e., alpha, a factor loading). In addition, we include a zero-inflation parameter z_j,

capturing excess zeros, and a shape parameter (ϕ_j; phi) for overdispersion, i.e., variance beyond what a Poisson count data model would imply. The intercept-slope parameterization results in

$$E(Y_{pj}) = (1 - z_j)\exp(\beta_j + \alpha_j\theta_p) \quad (1)$$

where $E(Y_{pj})$ is the expected count response for person p on item j, and θ_p is the level on the latent externalizing factor for person p. For $z_j = 0$, a log-linear model results, and the easiness is the log-expectation of an average person with $\theta_p = 0$. Values of $z_j > 0$ indicate that the item has excess zeros, i.e., z_j is the probability of the structural zero component. When $z_j > 0$ the relationship of the expected counts and the easiness is still log-linear, with $\ln(1 - z_j)$ as an offset. The item's discrimination parameter is how strongly the item is associated with the latent factor. In our study, easiness and discrimination provide information about the functioning and usefulness of each item and, collectively, the measure.

The resulting probability of observing a count y_{pj} in the zero-inflated negative binomial model is

$$P(Y_{pj} = y_{pj}) = \begin{cases} z_j + (1 - z_j)\left[\left(\frac{\phi_j}{\mu_{pj}+\phi_j}\right)^{\phi_j}\right] & if\ y_{pj} = 0 \\ (1 - z_j)\frac{\Gamma(y_{pj}+\phi_j)}{\Gamma(\phi_j)y_{pj}!}\left(\frac{\mu_{pj}}{\mu_{pj}+\phi_j}\right)^{y_{pj}}\left(\frac{\phi_j}{\mu_{pj}+\phi_j}\right)^{\phi_j} & if\ y_{pj} > 0 \end{cases} \quad (2)$$

where the mu (location) parameter, $\mu_{pj} = \exp(\beta_j + \alpha_j\theta_p)$, is the expected count in the negative binomial mixture component, i.e., $\ln(\mu_{pj}) = \beta_j + \alpha_j\theta_p$.

The Bayesian hierarchical model includes further assumptions: We estimated the probability of a structural zero, z_j, using a logistic mixed regression model while α_j, β_j, and θ_p each use a linear mixed model. We use the notation of the brms R-package (Bürkner, 2017), where the random effects of the form (… | item) all have mean zero normal distributions with freely estimated standard deviations. The person parameters' (theta) residual variance is fixed to

1, and the lower bound of the discrimination parameter is set to zero, following recommendations for model identification (Bürkner, 2020, 2021). We expected some curvilinearity given the wide age range (18–92 years of age) and the nonlinear trajectories of externalizing problems identified by prior studies (Bongers et al., 2004; Fanti & Henrich, 2010; Harris et al., 2025; Keijsers et al., 2012; Korhonen et al., 2018; Murray et al., 2022; Odgers et al., 2008; Petersen et al., 2015). We examined model fit using expected log predictive density (ELPD) from the widely applicable information criterion (WAIC). We compared a null model (with no predictors) with a linear, quadratic, and generalized additive model (GAM). The linear and quadratic models included the main effects of age and sex and their interaction. Below is the model formula for the GAM model.

$$\text{logit}(z_j) = 1 + (1|\text{item}) \tag{3}$$

$$\theta_p = \text{sex} + s(\text{age}, \text{sex}) + (1|\text{subject}) \tag{4}$$

$$\beta_j = 1 + (1|\text{item}) \tag{5}$$

$$\alpha_j = 1 + (1|\text{item}) \tag{6}$$

$$\log(\phi_i) = 1 + (1|\text{item}) \tag{7}$$

We modeled the sample's age-related course of latent externalizing problems by sex based on the estimated level on the latent externalizing problems (θ_p; theta) for each person p. Although there is a smooth interaction term in the model, i.e., $s(\text{age}, \text{sex})$, we also included the participant's sex as a fixed effect predictor of theta in Equation 4 for interpretability. We specified females as the reference group (female = 0, male = 1), because more females than males provided ratings. Here, $s(\ldots, \text{sex})$ is a sex-specific smoothing spline, i.e., the latent mean of externalizing problems (θ_p; theta) is a function of age and sex. This allows the latent level of externalizing problems to follow a nonlinear course across the wide age range (18–92 years of age). To prevent overfitting,

we use a *P*-spline-based smooth (i.e., penalized *B*-splines) using the s() function of the mgcv package 1.9-1 (Wood, 2017) in R 4.2.2 (R Core Team, 2022). For easier interpretation and identification, age in years was shifted (by subtracting 18) so that the value 0 represented an age of 18, the youngest age in the study.

In a Bayesian model, the final step is to specify prior distributions for all remaining parameters in the model. With the exception of the distributional assumptions for the item and person parameters discussed above, we kept the default priors used in the brms package (Bürkner, 2017), which uses vague but proper priors. The default priors for regression parameters are multivariate normal with mean zero and unknown covariance matrix Σ which follows a LKJ-correlation prior (Lewandowski et al., 2009). All standard deviation parameters are given a half Student-*t* prior with 3 degrees of freedom.

We fit the Bayesian longitudinal mixed models using the brms package 2.21.0 (Bürkner, 2017) in R, which uses the cmdstanr 0.8.0 (Gabry et al., 2024) interface to Stan 2.32.2 (Stan Development Team, 2020) for Bayesian modeling. In total, 8,000 post-warmup draws resulted from four chains with 4,000 iterations each, discarding 2,000 draws for warmup.

We identified the best-fitting form of age-related differences in the Bayesian IRT model. Model fit estimates are in Table S8. The GAM model fit the best. Thus, we selected the GAM model as our final model to allow curvilinearity in the age-related estimates, which thus provides better theory–model fit compared to the null and linear models. R-hat values (all less than 1.05) and trace plots demonstrated that the Bayesian IRT model converged. The posterior predictive check indicated that the model's predicted values were close to the observed values (see Figure S8).

We also intended to examine age and sex as predictors of the item parameters, easiness

and discrimination, as a test of differential item functioning (DIF). However, given the additional model complexity, the DIF models did not converge. Thus, we did not include age or sex as predictors of the item parameters; however, as noted above, we included age and sex as predictors of the person parameters (theta). We calculated item information from the model as described in Appendix S8. People's factor scores on the latent externalizing factor were estimated from the posterior distribution by averaging model-predicted posterior samples across chains and iterations, using the posterior_epred() function from the brms package.

Supplementary Appendix S8. Calculating Item Information from the Bayesian Item Response Theory Model.

The Fisher Information (FI) is used in item response theory models to find analytical expressions for the asymptotic standard error of person parameter estimates. The FI is model-specific, and the FI of neither the well-known Rasch Poisson Count Model (e.g., Doebler et al., 2014) nor the Conway Maxwell Poisson Counts Model (Beisemann, 2022) is appropriate for the negative binomial (NB) model with discrimination parameters proposed here. The model without discrimination parameters by Man and Harring (2019) results as a special case, so we also derive the Fisher Information for this model. We derive the FI in two steps: First, we study the NB case. In the second step, we generalize to the zero-inflated case.

Fisher Information of a Two-Parameter Negative Binomial Item

Omitting indices for ease of notation, the density of an NB random variable can be expressed as

$$f(n) = \text{NegBinomial2}(n|\mu, \phi) = \binom{n+\phi-1}{n}\left(\frac{\mu}{\mu+\phi}\right)^n\left(\frac{\phi}{\mu+\phi}\right)^\phi \tag{8}$$

with parameters $\mu = \exp(\alpha\theta + \beta) \in \mathbb{R}^+$ and $\phi \in \mathbb{R}^+$. This is called the alternative parametrization in the Stan language (Stan Development Team, 2020) and is a reparameterization of more common ways to denote the NB density. $S_\theta(n) = \frac{\partial}{\partial\theta}\ln\big(f(n)\big)$ denotes the score function, and because well-known mild regularity conditions hold, $I(\theta) = \mathrm{E}_\theta(S_\theta(n)^2)$ is the FI of a single NB-item. The log-likelihood function $\ell(n;\theta) := \ln\big(f(n)\big)$ is

$$\ell(n;\theta) = \ln p(n) = \ln\binom{n+\phi-1}{n} + n\ln\mu - (n+\phi)\ln(\mu+\phi) + \phi\ln\phi. \tag{9}$$

Note that $\ln\mu = \alpha\theta + \beta$. The derivative of $\ell(n;\theta)$ with respect to θ is

$$\frac{\partial}{\partial\theta}\ell(n;\theta) = n\frac{\partial}{\partial\theta}\ln\mu - (n+\phi)\frac{\partial}{\partial\theta}\ln(\mu+\phi). \tag{10}$$

Computing the partial derivatives with respect to μ and θ

$$\frac{\partial}{\partial\theta}\ln\mu = \alpha \text{ and } \frac{\partial}{\partial\theta}\ln(\mu+\phi) = \frac{1}{\mu+\phi}\frac{\partial\mu}{\partial\theta} = \frac{\mu\alpha}{\mu+\phi} \tag{12}$$

and plugging these in, we obtain

$$\frac{\partial}{\partial\theta}\ell(n;\theta) = \alpha\left[\frac{\phi(n-\mu)}{\mu+\phi}\right]. \tag{13}$$

As noted above, the FI $I(\theta)$ is the expected value of the squared derivative, so

$$I(\theta) = E\left[\left(\frac{\partial}{\partial\theta}\ell(n;\theta)\right)^2\right] = E\left[\left(\alpha\frac{\phi(n-\mu)}{\mu+\phi}\right)^2\right] = \left(\alpha\frac{\phi}{\mu+\phi}\right)^2 E[(n-\mu)^2] \tag{14}$$

Because n follows a negative binomial distribution with mean μ and variance

$$\mathrm{Var}[n] = E[(n-\mu)^2] = \mu + \frac{\mu^2}{\phi} \tag{15}$$

we can write

$$\begin{aligned} I(\theta) &= \left(\alpha\frac{\phi}{\mu+\phi}\right)^2\left(\mu+\frac{\mu^2}{\phi}\right) \\ &= \left(\alpha\frac{\phi}{\mu+\phi}\right)^2\left(\frac{\mu(\phi+\mu)}{\phi}\right) \\ &= \alpha^2\frac{\phi^2}{(\mu+\phi)^2}\cdot\frac{\mu(\mu+\phi)}{\phi} \\ &= \alpha^2\frac{\phi^2\mu(\mu+\phi)}{(\mu+\phi)^2\phi} \\ &= \alpha^2\frac{\phi\mu}{\mu+\phi} \end{aligned} \tag{16}$$

Zero-Inflated Case

In the zero-inflated case, the density of a single item is

$$\tilde{f}(n) = \begin{cases} \pi + (1-\pi)f(0) & n = 0 \\ (1-\pi)f(n) & n > 0, \end{cases} \tag{17}$$

where $0 \le \pi < 1$ is an additional parameter for zero-inflation. To derive the FI $\tilde{I}(\theta)$ for a single zero-inflated NB-item, we note that well-known mild regularity conditions hold and that we can use the representation

$$\tilde{I}(\theta) = \mathrm{E}\big(\tilde{S}_\theta(n)^2\big). \tag{18}$$

We now exploit that the zero-inflated NB is a discrete distribution. By re-arranging the infinite sums, we can express the FI with the help of the FI of the non-zero-inflated case (Equation 19):

$$\begin{aligned}
\tilde{I}(\theta) \quad &= \mathrm{E}\big(\tilde{S}_\theta(n)^2\big) = \sum_{n=0}^{\infty} \tilde{f}\,(n)\tilde{S}_\theta(n)^2 \\
&= \tilde{f}(0)\tilde{S}_\theta(0)^2 + \sum_{n=1}^{\infty} \tilde{f}_\theta\,(n)\tilde{S}_\theta(n)^2 \\
&= \big(\pi + (1-\pi)f(0)\big)\left(\frac{\partial}{\partial\theta}\ln\big(\pi + (1-\pi)f(0)\big)\right)^2 + \sum_{n=1}^{\infty}(1-\pi)\,f(n)\left(\frac{\partial}{\partial\theta}\ln\big((1-\pi)f(n)\big)\right)^2 \\
&= \big(\pi + (1-\pi)f(0)\big)\left(\frac{1}{\pi + (1-\pi)f(0)}(1-\pi)\frac{\partial}{\partial\theta}f(0)\right)^2 \\
&\quad + \sum_{n=1}^{\infty}(1-\pi)\,f(n)\left(\underbrace{\frac{\partial}{\partial\theta}\ln(1-\pi)}_{=\breve{0}} + \underbrace{\frac{\partial}{\partial\theta}\ln\big(f(n)\big)}_{=\breve{S_\theta}(n)}\right)^2 \\
&= \frac{(1-\pi)^2\left(\frac{\partial}{\partial\theta}f(0)\right)^2}{\pi + (1-\pi)f(0)} + \sum_{n=1}^{\infty}(1-\pi)\,f(n)S_\theta(n)^2 \\
&= \frac{(1-\pi)^2\left(\frac{\partial}{\partial\theta}f(0)\right)^2}{\pi + (1-\pi)f(0)} - (1-\pi)f(0)S_\theta(0)^2 + (1-\pi)\underbrace{\sum_{n=0}^{\infty} f\,(n)S_\theta(n)^2}_{=\breve{I}(\theta)} \\
&= \frac{(1-\pi)^2\left(\frac{\partial}{\partial\theta}f(0)\right)^2}{\pi + (1-\pi)f(0)} - (1-\pi)f(0)S_\theta(0)^2 + (1-\pi)I(\theta)
\end{aligned}$$

Where

$$S_\theta(0) = \tfrac{-\alpha\phi\mu}{\mu+\phi} \text{ and } I(\theta) = \tfrac{\alpha^2\phi\mu}{\mu+\phi} \tag{20}$$

and

$$\begin{aligned}\frac{\partial}{\partial\theta}f(0) &= \frac{\partial}{\partial\theta}\binom{0+\phi-1}{0}\left(\frac{\mu}{\mu+\phi}\right)^{0}\left(\frac{\phi}{\mu+\phi}\right)^{\phi} \\ &\qquad\qquad \overset{\smile}{=}1 \qquad\qquad \overset{\smile}{=}1 \\ &= \frac{\partial}{\partial\theta}\left(\frac{\phi}{\mu+\phi}\right)^{\phi} \\ &= \phi\left(\frac{\phi}{\phi+\mu}\right)^{\phi-1}\frac{0(\mu+\phi)-\phi\alpha\mu}{(\mu+\phi)^{2}} \\ &= -\frac{\phi^{\phi+1}\alpha\mu}{(\mu+\phi)^{\phi+1}}\end{aligned} \tag{21}$$

To implement this approach for estimating item information for items from the zero-inflated negative binomial model, we used the item_info_NB_zero_analytical() function of the petersenlab package version 1.1 (Petersen, 2025) in R version 4.3.1 (R Core Team, 2023). The item information functions are depicted in Supplementary Figure S12.

Test Information

We estimated test information for the set of items, at a given level of theta, as the sum of item information at that level of theta. The test information functions for the absolute frequency items, the dichotomized versions of the same items, and the items from the Adult Self-Report are in Figure 4.

Test Standard Error of Measurement

Test standard error of measurement was estimated as

$$\mathrm{SE}(\theta) = \frac{1}{\sqrt{I(\theta)}} \tag{22}$$

Test standard error of measurement for the absolute frequency items, the dichotomized versions of the same items, and the items from the Adult Self-Report are in Supplementary Figure S13.

Test Reliability

Test reliability was estimated as

$$\mathrm{reliability}(\theta) = \frac{I(\theta)}{I(\theta)+\sigma^{2}(\theta)} \tag{23}$$

In the GAM, this amounts to reliability conditional on age and sex. Test reliability for the absolute frequency items, the dichotomized versions of the same items, and the items from the Adult Self-Report are in Supplementary Figure S14.

References

Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA school-age forms & profiles*. University of Vermont, Research Center for Children, Youth, and Families.

Alexander, L. M., Salum, G. A., Swanson, J. M., & Milham, M. P. (2020). Measuring strengths and weaknesses in dimensional psychiatry. *Journal of Child Psychology and Psychiatry*, *61*(1), 40–50. https://doi.org/10.1111/jcpp.13104

Anderson-Butcher, D., Amorose, A. J., Iachini, A., & Ball, A. (2013). *Community and Youth Collaborative Institute School Experience Surveys*. College of Social Work, The Ohio State University.

Barkley, R. A. (2011). *Barkley Adult ADHD Rating Scale-IV (BAARS-IV)*. The Guilford Press.

Beisemann, M. (2022). A flexible approach to modelling over-, under- and equidispersed count data in IRT: The two-parameter Conway–Maxwell–Poisson model. *British Journal of Mathematical and Statistical Psychology*, *75*(3), 411–443. https://doi.org/10.1111/bmsp.12273

Bendixen, M., & Olweus, D. (1999). Measurement of antisocial behaviour in early adolescence and adolescence: Psychometric properties and substantive findings. *Criminal Behaviour and Mental Health*, *9*(4), 323–354. https://doi.org/10.1002/cbm.330

Bongers, I. L., Koot, H. M., Ende, J. v. d., & Verhulst, F. C. (2004). Developmental trajectories of externalizing behaviors in childhood and adolescence. *Child Development*, *75*(5), 1523–1537. https://doi.org/10.1111/j.1467-8624.2004.00755.x

Brown, B. B., Clasen, D. R., & Eicher, S. A. (1986). Perceptions of peer pressure, peer conformity dispositions, and self-reported behavior among adolescents. *Developmental Psychology*, *22*(4), 521–530. https://doi.org/10.1037/0012-1649.22.4.521

Brown, K., Atkins, M. S., Osborne, M. L., & Milnamow, M. (1996). A revised teacher rating scale for reactive and proactive aggression. *Journal of Abnormal Child Psychology*, *24*(4), 473–480. https://doi.org/10.1007/BF01441569

Budman, C. L., Rockmore, L., Stokes, J., & Sossin, M. (2003). Clinical phenomenology of episodic rage in children with Tourette syndrome. *Journal of Psychosomatic Research*, *55*(1), 59–65. https://doi.org/10.1016/S0022-3999(02)00584-6

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C. (2020). Analysing standard progressive matrices (SPM-LS) with Bayesian item response models. *Journal of Intelligence*, *8*(1), 5. https://doi.org/10.3390/jintelligence8010005

Bürkner, P.-C. (2021). Bayesian item response modeling in R with brms and Stan. *Journal of Statistical Software*, *100*(5), 1–54. https://doi.org/10.18637/jss.v100.i05

Burns, G. L., Lee, S., Servera, M., McBurnett, K., & Becker, S. P. (2018). *Child and Adolescent Behavior Inventory–Parent version 1.1*. https://static1.squarespace.com/static/5cdf50adf5b59400015725cd/t/5d5f4ce31064790001233ab2/1566526692691/CABI+Parent+1.1+2018+-+Complete+Measure.pdf

Burns, G. L., Walsh, J. A., Patterson, D. R., Holte, C. S., Sommers-Flanagan, R., & Parker, C. M. (2001). Attention deficit and disruptive behavior disorder symptoms: Usefulness of a frequency count rating procedure to measure these symptoms. *European Journal of Psychological Assessment*, *17*(1), 25–35. https://doi.org/10.1027//1015-5759.17.1.25

Carlson, G. A., Silver, J., & Klein, D. N. (2022). Psychometric properties of the Emotional Outburst Inventory (EMO-I): Rating what children do when they are irritable. *The Journal of Clinical Psychiatry*, *83*(2), 21m14015. https://doi.org/10.4088/JCP.21m14015

Carter, A. S., Briggs-Gowan, M. J., Jones, S. M., & Little, T. D. (2003). The Infant–Toddler Social and Emotional Assessment (ITSEA): Factor structure, reliability, and validity. *Journal of Abnormal Child Psychology*, *31*(5), 495–514. https://doi.org/10.1023/A:1025449031360

Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., Ader, D., Fries, J. F., Bruce, B., Rose, M., & on behalf of the Promis Cooperative Group. (2007). The Patient-Reported Outcomes Measurement Information System (PROMIS): Progress of an NIH roadmap cooperative group during its first two years. *Medical Care*, *45*(5), S3–S11. https://doi.org/10.1097/01.mlr.0000258615.42478.55

Choplin, E. G., Youngstrom, E., Hartsock, J. T., Smith, L. T., Anderson, S., Wilson, L., Howie, S., Salcedo, S., Thaker, I., Vincent, C., Kim, H., Van Meter, A., Langfus, J. A., Fogg, J. N., & Wall, A. (2018). *Child and Adolescent Disruptive Behavior Inventory (CADBI) screener*. https://osf.io/4j9gu

Conners, C. K. (2022). *Conners 4* (4th ed.). MHS.

Cook, C. R. (2012). *Student Externalizing Behavior Screener (SEBS)* APA PsycTests. https://doi.org/10.1037/t42705-000

Crick, N. R., & Grotpeter, J. K. (1995). Relational aggression, gender, and social-psychological adjustment. *Child Development*, *66*(3), 710–722. https://doi.org/10.1111/j.1467-8624.1995.tb00900.x

Curtiss, G., Rosenthal, R. H., Marohn, R. C., Ostrov, E., Offer, D., & Trujillo, J. (1983). Measuring delinquent behavior in inpatient treatment settings: Revision and validation of the Adolescent Antisocial Behavior Checklist. *Journal of the American Academy of Child Psychiatry*, *22*(5), 459–466. https://doi.org/10.1016/S0002-7138(09)61509-0

Daniels, B., Briesch, A. M., Volpe, R. J., & Owens, J. S. (2021). Content validation of direct behavior rating multi-item scales for assessing problem behaviors. *Journal of Emotional and Behavioral Disorders*, *29*(2), 71–82. https://doi.org/10.1177/1063426619882345

Doebler, A., Doebler, P., & Holling, H. (2014). A latent ability model for count data and application to processing speed. *Applied Psychological Measurement*, *38*(8), 587–598. https://doi.org/10.1177/0146621614543513

Egger, H. L., & Angold, A. (2004). The Preschool Age Psychiatric Assessment (PAPA): A structured parent interview for diagnosing psychiatric disorders in preschool children. In R. DelCarmen-Wiggins & A. Carter (Eds.), *Handbook of infant, toddler, and preschool mental health assessment* (pp. 223–243). Oxford University Press.

Elliott, D. S., & Ageton, S. S. (1980). Reconciling race and class differences in self-reported and official estimates of delinquency. *American Sociological Review*, *45*(1), 95–110. https://doi.org/10.2307/2095245

Erickson, M. L., & Empey, L. T. (1963). Court records, undetected delinquency and decision-making. *Journal of Criminal Law, Criminology and Police Science*, *54*, 456.

Essex, M. J., Boyce, W. T., Goldstein, L. H., Armstrong, J. M., Kraemer, H. C., & Kupfer, D. J. (2002). The confluence of mental, physical, social, and academic difficulties in middle childhood. II: Developing the MacArthur Health and Behavior Questionnaire. *Journal of*

the American Academy of Child & Adolescent Psychiatry, *41*(5), 588–603. https://doi.org/10.1097/00004583-200205000-00017

Eyberg, S. M., & Ross, A. W. (1978). Assessment of child behavior problems: The validation of a new inventory. *Journal of Clinical Child Psychology*, *7*(2), 113–116. https://doi.org/10.1080/15374417809532835

Fanti, K. A., & Henrich, C. C. (2010). Trajectories of pure and co-occurring internalizing and externalizing problems from age 2 to age 12: Findings from the National Institute of Child Health and Human Development Study of Early Child Care. *Developmental Psychology*, *46*(5), 1159–1175. https://doi.org/10.1037/a0020659

Farmer, C. A., & Aman, M. G. (2009). Development of the Children's Scale of Hostility and Aggression: Reactive/Proactive (C-SHARP). *Research in Developmental Disabilities*, *30*(6), 1155–1167. https://doi.org/10.1016/j.ridd.2009.03.001

Farrell, A. D., Sullivan, T. N., Goncy, E. A., & Le, A.-T. H. (2016). Assessment of adolescents' victimization, aggression, and problem behaviors: Evaluation of the Problem Behavior Frequency Scale. *Psychological Assessment*, *28*(6), 702–714. https://doi.org/10.1037/pas0000225

Frick, P. J. (2003). *The Inventory of Callous–Unemotional Traits* [Unpublished rating scale]. University of New Orleans.

Furlong, M. J., Smith, D. C., & Bates, M. P. (2002). Further development of the Multidimensional School Anger Inventory: Construct validation, extension to female adolescents, and preliminary norms. *Journal of Psychoeducational Assessment*, *20*(1), 46–65. https://doi.org/10.1177/073428290202000104

Gabry, J., Češnovar, R., Johnson, A., & Bronder, S. (2024). *cmdstanr: R Interface to CmdStan*. https://mc-stan.org/cmdstanr

Gadow, K. D., & Sprafkin, J. (1997). *Child Symptom Inventory–4 (CSI–4)*. Checkmate Plus.

Gartstein, M. A., & Rothbart, M. K. (2003). Studying infant temperament via the Revised Infant Behavior Questionnaire. *Infant Behavior and Development*, *26*(1), 64–86. https://doi.org/10.1016/S0163-6383(02)00169-8

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. *Journal of Child Psychology and Psychiatry*, *38*(5), 581–586. https://doi.org/10.1111/j.1469-7610.1997.tb01545.x

Gresham, F., & Elliott, S. N. (2008). *Social Skills Improvement System (SSIS) rating scales*. Pearson Assessments.

Halperin, J. M., McKay, K. E., & Newcorn, J. H. (2002). Development, reliability, and validity of the Children's Aggression Scale-Parent Version. *Journal of the American Academy of Child & Adolescent Psychiatry*, *41*(3), 245–252. https://doi.org/10.1097/00004583-200203000-00003

Harris, J. L., LeBeau, B., & Petersen, I. T. (2025). Reactive and control processes in the development of internalizing and externalizing problems across early childhood to adolescence. *Development and Psychopathology*, *37*(2), 836–858. https://doi.org/10.1017/S0954579424000713

Hindelang, M. J., Hirschi, T., & Weis, J. G. (1981). *Measuring delinquency*. Sage.

Hu, M.-C., Pavlicova, M., & Nunes, E. V. (2011). Zero-inflated and hurdle models of count data with extra zeros: Examples from an HIV-risk reduction intervention trial. *The American*

Journal of Drug and Alcohol Abuse, *37*(5), 367–375. https://doi.org/10.3109/00952990.2011.597280

Hubert, M., & Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. *Computational Statistics & Data Analysis*, *52*(12), 5186–5201. https://doi.org/10.1016/j.csda.2007.11.008

Huizinga, D. (1991). *Denver Youth Survey: Youth Interview Schedule*. University of Colorado-Boulder, Institute of Behavioral Science.

Inciardi, J. A. (1979). Heroin use and street crime. *Crime & Delinquency*, *25*(3), 335–346. https://doi.org/10.1177/001112877902500304

Jacobs, G. A., Phelps, M., & Rohrs, B. (1989). Assessment of anger expression in children: The pediatric anger expression scale. *Personality and Individual Differences*, *10*(1), 59–65. https://doi.org/10.1016/0191-8869(89)90178-5

Jellinek, M. S., Murphy, J. M., Robinson, J., Feins, A., Lamb, S., & Fenton, T. (1988). Pediatric Symptom Checklist: Screening school-age children for psychosocial dysfunction. *The Journal of Pediatrics*, *112*(2), 201–209. https://doi.org/10.1016/S0022-3476(88)80056-8

Johnson, B. D., Goldstein, P., Preble, E., Schmeidler, J., Lipton, D. S., Spunk, B., Duchaine, N., Norman, R., Miller, T., Meggert, N., Kale, A., & Hand, D. (1983). *Economic behavior of street opiate users. Final report to the National Institute on Drug Abuse and the National Institute of Justice.* Narcotic and Drug Research, Inc.

Kay, S. R., Wolkenfeld, F., & Murrill, L. M. (1988). Profiles of aggression among psychiatric patients: I. Nature and prevalence. *The Journal of Nervous and Mental Disease*, *176*(9), 539–546.

Kazdin, A. E., & Esveldt-Dawson, K. (1986). The interview for antisocial behavior: Psychometric characteristics and concurrent validity with child psychiatric inpatients. *Journal of Psychopathology and Behavioral Assessment*, *8*(4), 289–303. https://doi.org/10.1007/BF00960727

Keijsers, L., Loeber, R., Branje, S., & Meeus, W. (2012). Parent-child relationships of boys in different offending trajectories. A developmental perspective. *Journal of Child Psychology and Psychiatry*, *53*(12), 1222–1232. https://doi.org/10.1111/j.1469-7610.2012.02585.x

Kilgus, S. P., Eklund, K., von der Embse, N. P., Taylor, C. N., & Sims, W. A. (2016). Psychometric defensibility of the Social, Academic, and Emotional Behavior Risk Screener (SAEBRS) Teacher Rating Scale and multiple gating procedure within elementary and middle school samples. *Journal of School Psychology*, *58*, 21–39. https://doi.org/10.1016/j.jsp.2016.07.001

Korhonen, M., Luoma, I., Salmelin, R., Siirtola, A., & Puura, K. (2018). The trajectories of internalizing and externalizing problems from early childhood to adolescence and young adult outcome. *Journal of Child and Adolescent Psychiatry*, *2*(3), 7–12.

Koss, M. P., Abbey, A., Campbell, R., Cook, S., Norris, J., Testa, M., Ullman, S., West, C., & White, J. (2007). Revising the SES: A collaborative process to improve assessment of sexual aggression and victimization. *Psychology of Women Quarterly*, *31*(4), 357–370. https://doi.org/10.1111/j.1471-6402.2007.00385.x

Kwon, M., Lee, J.-Y., Won, W.-Y., Park, J.-W., Min, J.-A., Hahn, C., Gu, X., Choi, J.-H., & Kim, D.-J. (2013). Development and validation of a Smartphone Addiction Scale (SAS). *PloS One*, *8*(2), e56936. https://doi.org/10.1371/journal.pone.0056936

Ladd, G. W., & Profilet, S. M. (1996). The Child Behavior Scale: A teacher-report measure of young children's aggressive, withdrawn, and prosocial behaviors. *Developmental Psychology*, *32*(6), 1008–1024. https://doi.org/10.1037/0012-1649.32.6.1008

Landoll, R. R., La Greca, A. M., Lai, B. S., Chan, S. F., & Herge, W. M. (2015). Cyber victimization by peers: Prospective associations with adolescent social anxiety and depressive symptoms. *Journal of Adolescence*, *42*(1), 77–86. https://doi.org/10.1016/j.adolescence.2015.04.002

Lane, K. L., Oakes, W. P., Swogger, E. D., Schatschneider, C., Menzies, H. M., & Sanchez, J. (2015). Student Risk Screening Scale for internalizing and externalizing behaviors: Preliminary cut scores to support data-informed decision making. *Behavioral Disorders*, *40*(3), 159–170. https://doi.org/10.17988/0198-7429-40.3.159

Le Blanc, M., & Fréchette, M. (2013). *Male criminal activity from childhood through youth: Multilevel and developmental perspectives*. Springer.

LeBuffe, P. A., & Naglieri, J. A. (1999). *The Devereux early childhood assessment*. Kaplan Press.

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*(9), 1989–2001. https://doi.org/10.1016/j.jmva.2009.04.008

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, *83*(404), 1198–1202. https://doi.org/10.2307/2290157

Little, T. D., Henrich, C. C., Jones, S. M., & Hawley, P. H. (2003). Disentangling the "whys" from the "whats" of aggressive behaviour. *International Journal of Behavioral Development*, *27*(2), 122–133. https://doi.org/10.1080/01650250244000128

Loeber, R., Stouthamer-Loeber, M., Van Kammen, W. B., & Farrington, D. P. (1989). Development of a new measure of self-reported antisocial behavior for young children: Prevalence and reliability. In M. W. Klein (Ed.), *Cross-National Research in Self-Reported Crime and Delinquency* (Vol. 50, pp. 203–225). Springer Netherlands. https://doi.org/10.1007/978-94-009-1001-0_10

Luteijn, E., Jackson, S., Volkmar, F. R., & Minderaa, R. B. (1998). Brief Report: The development of the Children's Social Behavior Questionnaire: Preliminary data. *Journal of Autism and Developmental Disorders*, *28*(6), 559–565. https://doi.org/10.1023/A:1026060330122

Maechler, M., Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T., Koller, M., Conceicao, E. L., & di Palma, M. A. (2024). *robustbase: Basic robust statistics* (Version 0.99-2) [R package]. https://doi.org/10.32614/CRAN.package.robustbase

Man, K., & Harring, J. R. (2019). Negative binomial models for visual fixation counts on test items. *Educational and Psychological Measurement*, *79*(4), 617–635. https://doi.org/10.1177/0013164418824148

Marsee, M. A., & Frick, P. J. (2007). Exploring the cognitive and emotional correlates to proactive and reactive aggression in a sample of detained girls. *Journal of Abnormal Child Psychology*, *35*(6), 969–981. https://doi.org/10.1007/s10802-007-9147-y

McCarney, S. B., & Arthaud, T. J. (2001). *Emotional and Behavior Problem Scale–Second Edition*. Hawthorne Educational Services, Inc.

McCarney, S. B., & Arthaud, T. J. (2010). *Behavior Disorders Identification Scale: Renormed* (2nd ed.). Hawthorne Educational Services, Inc.

Merrell, K. W. (1996). Social-emotional assessment in early childhood: The Preschool and Kindergarten Behavior Scales. *Journal of Early Intervention*, *20*(2), 132–145. https://doi.org/10.1177/105381519602000205

Moffitt, T. E., & Silva, P. A. (1988). Self-reported delinquency: Results from an instrument for New Zealand. *Australian & New Zealand Journal of Criminology*, *21*(4), 227–240. https://doi.org/10.1177/000486588802100405

Murray, A. L., Nagin, D., Obsuth, I., Ribeaud, D., & Eisner, M. (2022). Young adulthood outcomes of joint mental health trajectories: A group-based trajectory model analysis of a 13-year longitudinal cohort study. *Child Psychiatry & Human Development*, *53*(5), 1083–1096. https://doi.org/10.1007/s10578-021-01193-8

Nadeau, J. M., McBride, N. M., Dane, B. F., Collier, A. B., Keene, A. C., Hacker, L. E., Cavitt, M. A., Alvaro, J. L., & Storch, E. A. (2016). Psychometric evaluation of the Rage Outbursts and Anger Rating Scale in an outpatient psychiatric sample. *Journal of Child and Family Studies*, *25*(4), 1229–1234. https://doi.org/10.1007/s10826-015-0303-7

Odgers, C. L., Moffitt, T. E., Broadbent, J. M., Dickson, N., Hancox, R. J., Harrington, H., Poulton, R., Sears, M. R., Thomson, W. M., & Caspi, A. (2008). Female and male antisocial trajectories: From childhood origins to adult outcomes. *Development and Psychopathology*, *20*(2), 673–716. https://doi.org/10.1017/S0954579408000333

Patton, J. H., Stanford, M. S., & Barratt, E. S. (1995). Factor structure of the Barratt Impulsiveness Scale. *Journal of Clinical Psychology*, *51*(6), 768–774. https://doi.org/10.1002/1097-4679(199511)51:6<768::AID-JCLP2270510607>3.0.CO;2-1

Pelham, W. E., Greiner, A. R., & Gnagy, E. M. (2008). *Student behavior teacher response observation code manual*.

Petersen, I. T. (2025). *petersenlab: A collection of R functions by the Petersen Lab* [R package]. https://doi.org/10.32614/CRAN.package.petersenlab

Petersen, I. T., Bates, J. E., Dodge, K. A., Lansford, J. E., & Pettit, G. S. (2015). Describing and predicting developmental profiles of externalizing problems from childhood to adulthood. *Development and Psychopathology*, *27*(3), 791–818. https://doi.org/10.1017/S0954579414000789

Peterson, M. A., Braiker, H. B., & Polich, R. (1980). *Doing crime: A survey of California prison inmates*. US Government Printing Office.

Peterson, M. A., Honig, P. K., Chaiken, J. M., & Ebener, P. A. (1982). *Survey of prison and jail inmates: Background and method*. RAND Corporation.

Plutchik, R., & van Praag, H. M. (1990). A self-report measure of violence risk, II. *Comprehensive Psychiatry*, *31*(5), 450–456. https://doi.org/10.1016/0010-440X(90)90031-M

Pontes, H. M., Király, O., Demetrovics, Z., & Griffiths, M. D. (2014). The conceptualisation and measurement of DSM-5 internet gaming disorder: The development of the IGD-20 test. *PloS One*, *9*(10), e110137. https://doi.org/10.1371/journal.pone.0110137

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

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org

Raine, A., Dodge, K., Loeber, R., Gatzke-Kopp, L., Lynam, D., Reynolds, C., Stouthamer-Loeber, M., & Liu, J. (2006). The reactive–proactive aggression questionnaire: Differential correlates of reactive and proactive aggression in adolescent boys. *Aggressive Behavior*, *32*(2), 159–171. https://doi.org/10.1002/ab.20115

Revelle, W. R. (2021). *psych: Procedures for personality and psychological research*. https://doi.org/10.32614/CRAN.package.psych

Reynolds, C. R., & Kamphaus, R. W. (2015). *Behavior Assessment System for Children–Third Edition (BASC–3)*. Pearson.

Rojahn, J., Matson, J. L., Lott, D., Esbensen, A. J., & Smalls, Y. (2001). The Behavior Problems Inventory: An instrument for the assessment of self-injury, stereotyped behavior, and aggression/destruction in individuals with developmental disabilities. *Journal of Autism and Developmental Disorders*, *31*(6), 577–588. https://doi.org/10.1023/A:1013299028321

Rothbart, M. K., Ahadi, S. A., Hershey, K. L., & Fisher, P. (2001). Investigations of temperament at three to seven years: The Children's Behavior Questionnaire. *Child Development*, *72*(5), 1394–1408. https://doi.org/10.1111/1467-8624.00355

Saunders, J. B., & Aasland, O. G. (1987). *WHO collaborative project on the identification and treatment of persons with harmful alcohol consumption. Report on phase I : The development of a screening instrument*. World Health Organization. https://iris.who.int/handle/10665/62031

Simms, L. J., Goldberg, L. R., Roberts, J. E., Watson, D., Welte, J., & Rotterman, J. H. (2011). Computerized adaptive assessment of personality disorder: Introducing the CAT–PD project. *Journal of Personality Assessment*, *93*(4), 380–389. https://doi.org/10.1080/00223891.2011.577475

Simonds, J., & Rothbart, M. K. (2004). *The Temperament in Middle Childhood Questionnaire (TMCQ): A computerized self-report instrument for ages 7–10* [Poster presentation]. Occasional Temperament Conference, Athens, GA, US.

Slocum-Gori, S. L., & Zumbo, B. D. (2011). Assessing the unidimensionality of psychological scales: Using multiple criteria from factor analysis. *Social Indicators Research*, *102*(3), 443–461. https://doi.org/10.1007/s11205-010-9682-8

Sobell, L. C., & Sobell, M. B. (1992). Timeline follow-back. In R. Z. Litten & J. P. Allen (Eds.), *Measuring alcohol consumption: Psychosocial and biochemical methods* (pp. 41–72). Humana Press. https://doi.org/10.1007/978-1-4612-0357-5_3

Sorgi, P., Ratey, J. J., Knoedler, D. W., Markert, R. J., & Reichman, M. (1991). Rating aggression in the clinical setting: A retrospective adaptation of the Overt Aggression

Scale: Preliminary results. *The Journal of Neuropsychiatry and Clinical Neurosciences*, *3*(2), S52–S56.

Spielberger, C. D., Jacobs, G., Russell, S., Crane, R. S., Jacobs, G. A., & Worden, T. J. (1985). The experience and expression of anger: Construction and validation of an anger expression scale. In M. A. Cheney & R. H. Rosenamn (Eds.), *Anger and hostility in cardiovascular behavioral disorder* (pp. 5–30). Hemisphere/McGraw-Hill.

Stan Development Team. (2020). *Stan modeling language users guide and reference manual*. https://mc-stan.org

Steele, R. G., Legerski, J.-P., Nelson, T. D., & Phipps, S. (2009). The Anger Expression Scale for Children: Initial validation among healthy children and children with cancer. *Journal of Pediatric Psychology*, *34*(1), 51–62. https://doi.org/10.1093/jpepsy/jsn054

Straus, M. A., Hamby, S. L., Boney-McCoy, S., & Sugarman, D. B. (1996). The Revised Conflict Tactics Scales (CTS2): Development and preliminary psychometric data. *Journal of Family Issues*, *17*(3), 283–316. https://doi.org/10.1177/019251396017003001

Stringaris, A., Goodman, R., Ferdinando, S., Razdan, V., Muhrer, E., Leibenluft, E., & Brotman, M. A. (2012). The Affective Reactivity Index: A concise irritability scale for clinical and research settings. *Journal of Child Psychology and Psychiatry*, *53*(11), 1109–1117. https://doi.org/10.1111/j.1469-7610.2012.02561.x

Sunderland, M., Slade, T., Krueger, R. F., Markon, K. E., Patrick, C. J., & Kramer, M. D. (2017). Efficiently measuring dimensions of the externalizing spectrum model: Development of the externalizing spectrum inventory-computerized adaptive test (ESI-CAT). *Psychological Assessment*, *29*(7), 868–880. https://doi.org/10.1037/pas0000384

Swanson, J. M., Deutsch, C., Cantwell, D., Posner, M., Kennedy, J. L., Barr, C. L., Moyzis, R., Schuck, S., Flodman, P., Spence, M. A., & Wasdell, M. (2001). Genes and attention-deficit hyperactivity disorder. *Clinical Neuroscience Research*, *1*(3), 207–216. https://doi.org/10.1016/S1566-2772(01)00007-X

Swanson, J. M., Kraemer, H. C., Hinshaw, S. P., Arnold, L. E., Conners, C. K., Abikoff, H. B., Clevenger, W., Davies, M., Elliott, G. R., Greenhill, L. L., Hechtman, L., Hoza, B., Jensen, P. S., March, J. S., Newcorn, J. H., Owens, E. B., Pelham, W. E., Schiller, E., Severe, J. B., . . . 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 Academy of Child & Adolescent Psychiatry*, *40*(2), 168–179. https://doi.org/10.1097/00004583-200102000-00011

Thalmayer, A. G., Marshall, J., & Scalise, K. (2023). The International Mental Health Assessment: Validation of an efficient screening inventory. *Collabra: Psychology*, *9*(1). https://doi.org/10.1525/collabra.74546

Thornberry, T. P., & Farnworth, M. (1982). Social correlates of criminal involvement: Further evidence on the relationship between social status and criminal behavior. *American Sociological Review*, *47*(4), 505–518. https://doi.org/10.2307/2095195

Tierney, N., & Cook, D. (2023). Expanding tidy data principles to facilitate missing data exploration, visualization and assessment of imputations. *Journal of Statistical Software*, *105*(7), 1–31. https://doi.org/10.18637/jss.v105.i07

Wakschlag, L. S., Briggs-Gowan, M. J., Choi, S. W., Nichols, S. R., Kestler, J., Burns, J. L., Carter, A. S., & Henry, D. (2014). Advancing a multidimensional, developmental spectrum approach to preschool disruptive behavior. *Journal of the American Academy of*

Child & Adolescent Psychiatry, *53*(1), 82–96.e83. https://doi.org/10.1016/j.jaac.2013.10.011

Waschbusch, D. A., & Elgar, F. J. (2007). Development and validation of the conduct disorder rating scale. *Assessment*, *14*(1), 65–74. https://doi.org/10.1177/1073191106289908

Watkins, L. E., Maldonado, R. C., & DiLillo, D. (2018). The Cyber Aggression in Relationships Scale: A new multidimensional measure of technology-based intimate partner aggression. *Assessment*, *25*(5), 608–626. https://doi.org/10.1177/1073191116665696

West, D. J., & Farrington, D. P. (1977). *The delinquent way of life: Third report of the Cambridge Study in Delinquent Development*. Heinemann Educational Books.

White, H. R., & Labouvie, E. W. (1989). Towards the assessment of adolescent problem drinking. *Journal of Studies on Alcohol*, *50*(1), 30–37. https://doi.org/10.15288/jsa.1989.50.30

Willoughby, M., Kupersmidt, J., & Bryant, D. (2001). Overt and covert dimensions of antisocial behavior in early childhood. *Journal of Abnormal Child Psychology*, *29*(3), 177–187. https://doi.org/10.1023/A:1010377329840

Woicik, P. A., Stewart, S. H., Pihl, R. O., & Conrod, P. J. (2009). The substance use risk profile scale: A scale measuring traits linked to reinforcement-specific substance use profiles. *Addictive Behaviors*, *34*(12), 1042–1055. https://doi.org/10.1016/j.addbeh.2009.07.001

Wolraich, M. L., Feurer, I. D., Hannah, J. N., Baumgaertel, A., & Pinnock, T. Y. (1998). Obtaining systematic teacher reports of disruptive behavior disorders utilizing DSM-IV. *Journal of Abnormal Child Psychology*, *26*(2), 141–152. https://doi.org/10.1023/A:1022673906401

Wood, S. N. (2017). *Generalized additive models: An introduction with R* (2nd ed.). CRC press. https://doi.org/10.1201/9781315370279

Woods-Groves, S. (2015). The Human Behavior Rating Scale–Brief: A tool to measure 21 century skills of K–12 learners. *Psychological Reports*, *116*(3), 769–796. https://doi.org/10.2466/03.11.PR0.116k29w0

Young, K. S. (1998). *Caught in the net: How to recognize the signs of internet addiction—and a winning strategy for recovery*. John Wiley & Sons, Inc.

Zeman, J., Shipman, K., & Suveg, C. (2002). Anger and sadness regulation: Predictions to internalizing and externalizing symptoms in children. *Journal of Clinical Child & Adolescent Psychology*, *31*(3), 393–398. https://doi.org/10.1207/S15374424JCCP3103_11

Supplementary Table S1.

Example Measures of Externalizing Behavior (and Related Constructs) That Use Vague Quantifiers of Relative Frequency (or Ability/Recency/Severity).

Type	Scale Points	Response Options	Instrument (and Citation)
ability/disability	7	far above average; above average; slightly above average; average; slightly below average; below average; far below average	Strengths and Weaknesses of Attention-Deficit/Hyperactivity-symptoms and Normal-behaviors (SWAN; Swanson, Deutsch, et al., 2001)
ability/disability	7	far above average; above average; slightly above average; average; slightly below average; below average; far below average	Extended Strengths and Weaknesses of Attention-Deficit/Hyperactivity-symptoms and Normal-behaviors (E–SWAN; Alexander et al., 2020)
agreement	3	not true; somewhat true; certainly true	Strengths and Difficulties Questionnaire (SDQ; Goodman, 1997)
agreement	3	not true; somewhat true; certainly true	Affective Reactivity Index (ARI; Stringaris et al., 2012)
agreement	4	not at all; just a little; quite a bit; very much	Swanson, Nolan, and Pelham Rating Scale–IV (SNAP–IV; Swanson, Kraemer, et al., 2001)
agreement	4	not at all; just a little; pretty much; very much	Adolescent Antisocial Behavior Checklist (AABC; Curtiss et al., 1983)
agreement	4	not true at all about me; 2 midpoint anchors; completely true about me	Little's Forms and Functions of Aggression (Little et al., 2003)
agreement	4	strongly disagree; disagree; agree; strongly agree	Substance Use Risk Profile Scale (SURPS; Woicik et al., 2009)
agreement	4	not at all true; somewhat true; very true; definitely true	Inventory of Callous-Unemotional Traits (ICU; Frick, 2003)

agreement	4	not true at all; just a little; pretty much true; very much true	Conners–4 (Conners, 2022)
agreement	4	not at all true; somewhat true; very true; definitely true	Peer Conflict Scale (PCS; Marsee & Frick, 2007)
agreement	5	The child has never exhibited this behavior or attribute; (Use this number when the choice between '1' and '3' is difficult); The child displays this behavior or attribute about the same as other children his or her age; (Use this number when the choice between '3' and '5' is difficult); The child displays this behavior or attribute to a very high degree	Human Behavior Rating Scale (HBRS; Woods-Groves, 2015)
agreement	5	strongly disagree; disagree; neither agree or disagree; agree; strongly agree	Internet Gaming Disorder Test (IGD–20; Pontes et al., 2014)
agreement	5	very untrue of me; moderately untrue of me; neither true nor untrue of me; moderately true of me; very true of me	Comprehensive Assessment of Traits Relevant to Personality Disorder (CAT–PD; Simms et al., 2011)
agreement	6	strongly disagree; disagree; weakly disagree; weakly agree; agree; strongly agree	Smartphone Addiction Scale (SAS; Kwon et al., 2013)
agreement	7	extremely untrue of your child; quite untrue of your child; slightly untrue of your child; neither true nor false of your child; slightly true of your child; quite true of your child; extremely true of your child	Children's Behavior Questionnaire (CBQ; Rothbart et al., 2001)
frequency	3	hardly ever; sometimes; often	Pediatric Anger Expression Scale (PAES; Jacobs et al., 1989)
frequency	3	never; sometimes; often	Pediatric Symptom Checklist (PSC; Jellinek et al., 1988)
frequency	3	never; sometimes; often	Reactive–Proactive Aggression Questionnaire (RPAQ; Raine et al., 2006)
frequency	3	not true; somewhat true/sometimes; very true/often	Infant–Toddler Social Emotional Assessment (ITSEA; Carter et al., 2003)

frequency	3	never; sometimes; often	Children's Social Behavior Questionnaire (CSBQ; Luteijn et al., 1998)
frequency	3	doesn't apply (seldom displays this behavior); applies sometimes (occasionally displays this behavior); certainly applies (often displays this behavior)	Child Behavior Scale (CBS; Ladd & Profilet, 1996)
frequency	3	often true; sometimes true; not true	Behavior Problems Inventory (BPI; Rojahn et al., 2001)
frequency	3	hardly ever; sometimes; often	Children's Anger Management Scales (CAMS; Zeman et al., 2002)
frequency	3	never; sometimes; very often	Revised Teacher Rating Scale for Reactive and Proactive Aggression (RPRA; Brown et al., 1996)
frequency	4	never; occasionally (sometimes); often; always	Multidimensional School Anger Inventory (MSAI; Furlong et al., 2002)
frequency	4	never; rarely/seldom; occasionally; frequently	Student Externalizing Behavior Screener (SEBS; Cook, 2012)
frequency	4	never; occasionally; often; very often	Vanderbilt ADHD Diagnostic Parent/Teacher Rating Scale (Wolraich et al., 1998)
frequency	4	never or rarely; sometimes; often; very often	Barkley Adult ADHD Rating Scale (BAARS; Barkley, 2011)
frequency	4	never; sometimes; often; almost always	Behavior Assessment System for Children–Third Edition (BASC–3; Reynolds & Kamphaus, 2015)
frequency	4	almost never; sometimes; often; almost always	Anger Expression Scale (AES; Spielberger et al., 1985); Anger Expression Scale for Children (AESC; Steele et al., 2009)
frequency	4	never; sometimes; often; almost always	Social Skills Improvement System (SSIS; Gresham & Elliott, 2008)

frequency	4	never; occasionally; sometimes; frequently	Student Risk Screening Scale–Internalizing & Externalizing (SRSS–IE; Lane et al., 2015)
frequency	4	never; sometimes; often; very often	Child/Adolescent Symptom Inventory–4 (CSI/ASI–4; Gadow & Sprafkin, 1997)
frequency	4	never; sometimes; often; almost always	Social, Academic, and Emotional Behavior Risk Screener (SAEBRS; Kilgus et al., 2016)
frequency	4	never; rarely true; sometimes true; often true	Preschool and Kindergarten Behavior Scales (PKBS; Merrell, 1996)
frequency	4	never; sometimes; often; very often	Past Feelings and Acts of Violence (PFAV; Plutchik & van Praag, 1990)
frequency	5	never; almost never; sometimes; almost all of the time; all of the time	Children's Social Behavior Scale (CSBS; Crick & Grotpeter, 1995)
frequency	5	never; almost never; sometimes; almost always; always	Patient Reported Outcome Measurement Information System (PROMIS; Cella et al., 2007)
frequency	5	not at all; rarely; occasionally; often; always	Internet Addiction Test (IAT; Young, 1998)
frequency	5	never; rarely; occasionally; frequently; very frequently	Devereux Early Child Assessment (DECA; LeBuffe & Naglieri, 1999)
frequency	5	very often; often; sometimes; seldom; never	Community and Youth Collaborative Institute School Experience Surveys (CAYCI; Anderson-Butcher et al., 2013)
frequency	6	rarely; never; occasionally; often; almost always; always	Barratt Impulsiveness Scale (BIS; Patton et al., 1995)
frequency	7	never; very rarely; less than half the time; about half the time; almost always; always	Infant Behavior Questionnaire–Revised (IBQ–R; Gartstein & Rothbart, 2003)
frequency	7	never; seldom; sometimes; often; always	Eyberg Child Behavior Inventory (ECBI; Eyberg & Ross, 1978)

hybrid	3	never or not true; sometimes or somewhat true; often or very true	MacArthur Health and Behavior Questionnaire (HBQ; Essex et al., 2002)
hybrid	3	not true; somewhat or sometimes true; very true or often true	Child Behavior Checklist (CBCL; Achenbach & Rescorla, 2001) and Achenbach System of Empirically Based Assessment (ASEBA)
hybrid	3	does not happen; mild or infrequent problem; moderately serious and/or frequent problem	Children's Scale of Hostility and Aggression: Reactive/Proactive (C-SHARP; Farmer & Aman, 2009)
hybrid	5	almost always untrue; usually untrue; sometimes true or sometimes untrue; usually true; almost always true	Temperament in Middle Childhood Questionnaire (TMCQ; Simonds & Rothbart, 2004)
recency	3	recent or new problem (6 months or less); long time (more than 6 months); always	Interview for Antisocial Behavior (IAB; Kazdin & Esveldt-Dawson, 1986)
severity (expanded format)	5	response options are particular behaviors ranging from less to more severe; includes labels such as "minor", "major", "serious"	Modified Overt Aggression Scale (MOAS; Kay et al., 1988)
severity	6	not a problem; slight problem; moderate problem; serious problem (responses allow midpoints)	Direct Behavior Rating Multi-Item Scales–Externalizing (DBR–MIS–E; Daniels et al., 2021)

Supplementary Table S2.

Relative Strengths and Weaknesses of Vague and Numeric Quantifiers.

Consideration	Vague Quantifiers	Closed-Ended Numeric Quantifiers	Open-Ended Numeric Quantifiers
Time to complete	shorter	shorter	longer
Ease of completion	easier	easier	harder (but can be reduced by allowing participants to select the timeframe on which it is easiest to estimate the frequency of a behavior)
Potential for item-level missingness	lower	lower	higher
Difficulty to score	lower	lower	higher/may require more complex scoring approaches, such as latent variable modeling that require larger samples
Measurement error as behavior frequency increases	may increase	may increase	may increase
Subjectivity/objectivity	more subjective	more objective	more objective
Meaning of quantifier…			
by rater (e.g., attitude toward behavior; their frequency of engaging in behavior)	differs	presumably does not differ	presumably does not differ
by demographic characteristics of rater	differs	presumably does not differ	presumably does not differ
by culture	differs	presumably does not differ	presumably does not differ
by situation	differs	presumably does not differ	presumably does not differ
by context in which quantifier is used (e.g., the response options and their order)	differs	presumably does not differ	presumably does not differ
by item/behavior/event of interest (e.g., base rate)	differs	presumably does not differ	presumably does not differ
Susceptibility to bias including cultural bias	higher	lower	lower

Conflates frequency and impairment	yes	no	no
Involves social referencing (i.e., reference group)	yes	no	no
Reference group used differs between raters	yes	n/a	n/a
Relation between actual frequency and reported frequency	can be inversely related	presumably positively related	presumably positively related
Different items may require different frequency response options	no	yes	no
Provides normative presumptions to the respondent for how commonly the behavior tends to occur in the population, which may elicit lower endorsement rates for sensitive/stigmatized behavior	no	yes	no
Frequency ranges provided by the researchers alter responses	n/a	yes	no
Artificially restricted number of response options	yes	yes	no
Potential for restricted variability	higher	higher	lower
Potential for sensitivity to change (responsivity)	lower	lower	higher
Potential for sensitivity to differences in level across development or groups	lower; can obscure differences across development/groups due to social referencing	medium	higher
Distance between the response categories	differs	can differ	does not differ
Capacity to assess the full range of individual differences	lower	medium	higher
Potential for examining associations with external criteria	lower	medium	higher
Potential for understanding construct of interest	lower	medium	higher
Potential for understanding how frequent a behavior is considered developmentally typical or atypical at a given age	lower	medium	higher
Precision	lower	medium	higher

Interpretability	lower	medium	higher
Meaningfulness of central tendency	lower	medium	higher

Note. Gray cells indicate relative strengths. Rows in the top section (i.e., before the first dashed line) indicate relative strengths of vague quantifiers. Rows in the middle section (i.e., in between the two dashed lines) indicate shared weaknesses of all quantifiers. Rows in the bottom section (i.e., after the second dashed line) indicate relative strengths of numeric quantifiers.

Supplementary Table S3.

Example Measures of Externalizing Behavior (and Related Constructs) That Use Numeric Quantifiers of Absolute Frequency.

Quantifier	Scale Points	Response Options	Reference Timeframe	Instrument (and Citation)
numeric	3	never; once or twice; three or more times	not specified, implied lifetime	Self-Report Early Delinquency (SRED; Moffitt & Silva, 1988)
vague/numeric hybrid	3	never; once or twice; more often	past 6 months	Self-Report of Antisocial Behavior Scale (SRABS; Loeber et al., 1989)
vague/numeric hybrid	3	never; once or twice; many times or often	past year	Self-Administered Questionnaire of Delinquency (SAQD; Le Blanc & Fréchette, 2013)
numeric	4	0; 1; 2; 3+	past 12 months; since age 14	Sexual Experiences Survey–Perpetration (SES–P; Koss et al., 2007)
numeric	4	never; once; two or three times; four or more (the specific numeric ranges differ from item to item)	past 3 years	Self-Reported Delinquency (West & Farrington, 1977)
numeric	5	never; once; monthly; weekly; daily	past 12 months	Conduct Disorder Rating Scale (CDRS; Waschbusch & Elgar, 2007)
numeric	5	never; 1–2 times; 3–5 times; 6–10 times; more than 10 times	past year	Rutgers Alcohol Problem Index (RAPI; White & Labouvie, 1989)
numeric	5	never; once a month or less; once a week or less; 2–3 times a week; most days; some questions are: once or twice; 3–5 times; 6–10 times; more than 10 times	past year	Children's Aggression Scale (CAS; Halperin et al., 2002)
numeric	5	never; once or twice; a few times; about once a week; a few times a week	not specified	(Revised) Peer Experiences Questionnaire (R–PEQ; Prinstein et al., 2001)

numeric	5	never; once or twice; a few times; about once a week; a few times a week	past 2 months	Cyber Peer Experiences Questionnaire (C–PEQ; Landoll et al., 2015)
numeric	5	once/month or less; once/week; three to four times/week; once/day; many times/day	not specified	Types of Antisocial Behavior (TAB; Willoughby et al., 2001)
numeric	5	never; monthly or less; 2–4 times a month; 2–3 times a week; 4 or more times a week	some items, but not all, specify past year	Alcohol Use Disorders Identification Test (AUDIT; Saunders & Aasland, 1987)
numeric	5	0 times; 1–2 times; 3–5 times; 6–10 times; more than 10 times; if more than 10, participant is asked the number of times in an open-ended way	the 3 years prior to beginning of prison term	Prison Survey Questionnaire; (PSQ; Peterson et al., 1980)
vague/numeric hybrid	5	never (0 times); sometimes (1–2 times); often (3–4 times); usually (5–10 times); always (> 10 times)	past week	Retrospective Overt Aggression Scale (ROAS; Sorgi et al., 1991)
vague/numeric hybrid	5	never; once or twice; 3 or 4 times; pretty often or almost every day (adaptation by Posner & Vandell: never; once or twice; about once a week; 2–3 times per week; 4 or more times a week)	past month (adaptation by Posner & Vandell: since beginning of school term)	Self-Reported Behavior Index (SRBI; Brown et al., 1986), adapted as the Misconduct Scale (MS) by Posner & Vandell
numeric	6	never; 1–2 times; 3–5 times; 6–9 times; 10–19 times; 20 or more times	past 30 days	Problem Behavior Frequency Scale (PBFS; Farrell et al., 2016)
vague/numeric hybrid	6	never; rarely (less than once per week); some (1–3) days of the week; most (4–6) days of the week; every day of the week; many times each day	past month	Multidimensional Assessment of Preschool Disruptive Behavior (MAP–DB; Wakschlag et al., 2014)
vague/numeric hybrid	6	almost never (never or about once per month); seldom (about once per week); sometimes (several times per week); often (about once per day); very	past month	Child and Adolescent Behavior Inventory (CABI; Burns et al., 2018)

		often (several times per day); almost always (many times per day)		
vague/numeric hybrid	6	never; rarely; several times a month; weekly; at least 3 times/week; daily	not specified	Emotional Outburst Inventory (EMO–I; Carlson et al., 2022)
numeric	7	not in the last month (or never); once a month; about twice a month; once a week; about twice a week; more than twice a week / half the days; daily or almost daily	past month	International Mental Health Assessment (IMHA; Thalmayer et al., 2023)
numeric	7	not in my presence; one time in several months; several times, up to one time a month; more than one time a month, up to one time a week; more than one time a week, up to once a day; more than once a day, up to once an hour; more than once an hour	not specified	Emotional and Behavior Problem Scale–2 (EBPS–2; McCarney & Arthaud, 2001)
numeric	7	not in my presence; one time in several months; several times, up to one time a month; more than one time a month, up to one time a week; more than one time a week, up to once a day; more than once a day, up to once an hour; more than once and hour	not specified	Behavior Disorders Identification Scale (BDIRS–2:R; McCarney & Arthaud, 2010)
numeric	7	not in the past 6 months; one time in the past 6 months; two times in the past 6 months; once per month in the past 6 months; once per week in the past 6 months; once per day in the past 6 months; more than once per day in the past 6 months	past 6 months	Child and Adolescent Disruptive Behavior Inventory–Parent (CADBI–P; Burns et al., 2001)
numeric	7	never; once a month; once every 2–3 weeks; once a week; 2–3 times a week; once a day; 2–3 times a day	past year	Self-Report Delinquency (SRD) in the National Youth Survey (Elliott & Ageton, 1980)
numeric	8	never in past month; 1–2 times in past month; 3-4 times in past month; 2–6 times per week; 1 time per day; 2–5 times per day; 6–9 times per day; 10 or more times per day	past month	Child and Adolescent Disruptive Behavior Inventory Screener (CADBI–S; Choplin et al., 2018)

numeric	8	this has never happened; once in the past 6 months; twice in the past 6 months; 3–5 times in the past 6 months; 6–10 times in the past 6 months; 11–20 times in the past 6 months; more than 20 times in the past 6 months; not in the past 6 months, but it did happen before	past 6 months	Cyber Aggression in Relationships Scale (CARS; Watkins et al., 2018)
numeric	8	this has never happened; not in the past year, but it did happen before; once in the past year; twice in the past year; 3–5 times in the past year; 6–10 times in the past year; 11–20 times in the past year; more than 20 times in the past year	past year	Revised Conflict Tactics Scale (CTS2; Straus et al., 1996)
numeric	unknown		as a juvenile (before age 18); as an adult (since age 18)	Self-Report of Delinquency and Crime (SRDC; Thornberry & Farnworth, 1982)
numeric	open-ended	number of drinking days; number of drinks consumed on every drinking day	varies; ranges from 7 days to 24 months	Timeline Follow-back (TLFB; Sobell & Sobell, 1992)
numeric	open-ended	observer tally of child behaviors during a 30-minute interval	30 minutes	Student Behavior Teacher Response System (SBTR; Pelham et al., 2008)
numeric	open-ended	First, the participant endorsed whether they have ever done/taken part in a behavior; if they indicated "Yes", they were asked how many times they engaged in the behavior: "about how many times this spring (from Christmas until now)?"	academic term—i.e., "this spring (from Christmas until now)"; ~ 5 months	Bergen Questionnaire on Antisocial Behavior (BQAB; Bendixen & Olweus, 1999)
numeric	open-ended	number of instances per day for each of various criminal behaviors that resulted in economic benefit	past 7 days	Economic Behavior Study Weekly/Daily Data Collection Form (EBS; Johnson et al., 1983)

numeric	open-ended	"How many times in the last year have you…"?; "How often in the last year have you used…"?	past year	Self-Report Delinquency (SRD) in the National Youth Survey (Elliott & Ageton, 1980)
numeric	open-ended	number of instances for each of various delinquent behaviors (up to a maximum of 99 per behavior)	past year	Seattle Self-Report Instrument (SSRI; Hindelang et al., 1981)
numeric	open-ended	number of instances for each of various criminal behaviors	past year	Current Criminal Activity (CCA; Inciardi, 1979)
numeric	open-ended	number of instances for each of various criminal behaviors	cumulative (i.e., during adolescence and up to the mid-20s) and, separately, per year	Semi-structured Interview of Delinquency (SSID; Le Blanc & Fréchette, 2013)
numeric	open-ended	number of instances for each of various criminal behaviors	lifetime	Interview of Offending (Erickson & Empey, 1963)
vague/numeric hybrid	open-ended	First, the participant endorsed whether they have ever done/taken part in a behavior; if they indicated "Yes", they indicated how many times they engaged in the behavior: "1 to 10" or "11 or more". If they indicated 1 to 10, they were asked how many times. If they indicated "11 or more", they were asked how many months they engaged in the behavior and the most appropriate timeframe ("everyday or almost everyday"; "several times a week; "every week or almost every week"; "less than every week") and how many times per that timeframe (e.g., per day, per week, or per month).	January 1 of the year preceding their arrest for the current conviction through the month of their arrest; 13–24 months	Rand Corporation Survey of Prison and Jail Inmates (Peterson et al., 1982)

Measures with only one or a few numeric quantifier items

numeric	4	no rage attacks occurred; 1–2 rage attacks occurred; 3–7 rage attacks occurred, but not every day; at least one rage attack occurred each day	past 7 days	Rage Outbursts and Anger Rating Scale (ROARS; Nadeau et al., 2016)
numeric	open-ended	"How many times must you tell X to do something before s/he will do it?"	not specified	Preschool Age Psychiatric Assessment (PAPA; Egger & Angold, 2004)
numeric	open-ended	"How many rage attacks of this type has your child had over the past week?"	past week	Rage Attacks Questionnaire (RAQ; Budman et al., 2003)
numeric	open-ended	"How many times in the past year?"	past year	Denver Youth Survey (DYS; Huizinga, 1991)

Supplementary Table S4.

Racial/Ethnic Composition of the Sample.

Race/Ethnicity	*n*	%	Subclassification	*n*	%
Hispanic, Latino/a/x, or Spanish origin	86	7.0			
			Mexican, Mexican American, or Chicano	47	3.8
			Puerto Rican	12	1.0
			Cuban	0	0.0
			Another Hispanic, Latino/a/x, or Spanish origin	27	2.2
White	1,137	92.0		1,137	92.0
Black or African American	44	3.6		44	3.6
American Indian or Alaska Native	15	1.2		15	1.2
Asian	64	5.2			
			Asian Indian	9	0.7
			Chinese	15	1.2
			Filipino	14	1.1
			Japanese	5	0.4
			Korean	4	0.3
			Vietnamese	11	0.9
			Other Asian	10	0.8
Native Hawaiian or Other Pacific Islander	1	0.1			
			Native Hawaiian	1	0.1
			Guamanian or Chamorro	0	0.0
			Samoan	0	0.0
			Other Pacific Islander	0	0.0
Multiracial	55	4.4		55	4.4
Other Race	31	2.5			
			Arab, Middle Eastern, or North African	11	0.9
			Other	20	1.6

Note. The racial and ethnic categories specified are based on the 2020 U.S. Census. We added an option for "Arab, Middle Eastern, or North African." The cells sum to greater than 100% because participants were allowed to select multiple racial categories.

Supplementary Table S5.

Externalizing Problem Items

Variable	Item	Wording
es_frequency_1	1	I show off
es_frequency_2	2	I brag or boast
es_frequency_6	6	I feel like I cannot stop engaging in an unessential activity
es_frequency_7	7	I blame others for my own mistakes
es_frequency_8	8	I blame others for my poor performance
es_frequency_10	10	I get bored during a given activity
es_frequency_14	14	I injure other people to see them in pain
es_frequency_15	15	I threaten others
es_frequency_16	16	I hurt others' feelings on purpose
es_frequency_17	17	I have fights with others to show who is on top
es_frequency_18	18	I use physical force, or threaten to use force, to get others to do what I want
es_frequency_27	27	I tell little lies or "white" lies
es_frequency_28	28	I tell big lies
es_frequency_29	29	I cheat in games or activities
es_frequency_30	30	I lie to get out of things that I don't want to do
es_frequency_31	31	I lie to get ahead at work or school
es_frequency_32	32	I talk someone into having sex with me
es_frequency_33	33	I get things from people by making them feel sorry for me
es_frequency_34	34	I exaggerate when describing events
es_frequency_39	39	I miss things I promised to attend
es_frequency_40	40	I have problems at work or school because I was irresponsible
es_frequency_41	41	I complete tasks or assignments late
es_frequency_42	42	I do not live up to my end of a contract
es_frequency_51	51	I break or destroy things belonging to others on purpose

es_frequency_52	52	I vandalize public property
es_frequency_53	53	I vandalize a house or private property
es_frequency_54	54	I damage property
es_frequency_58	58	I hang around people who get in trouble
es_frequency_59	59	My friends get drunk
es_frequency_60	60	My friends use drugs
es_frequency_61	61	My friends commit violent or property crime
es_frequency_62	62	My friends get suspended or fired
es_frequency_64	64	I avoid or am reluctant to engage in tasks that require sustained mental effort
es_frequency_65	65	I lose things
es_frequency_70	70	I try to intimidate others
es_frequency_71	71	I send messages to someone after they asked me to stop
es_frequency_72	72	I check or track someone's Internet activity without their permission
es_frequency_76	76	I get hurt doing something for the thrill
es_frequency_77	77	I do something risky for enjoyment, fun, or pleasure
es_frequency_78	78	I do something fun that others think is too dangerous
es_frequency_79	79	I do something illegal because it is exciting
es_frequency_80	80	I seek out dangerous situations
es_frequency_81	81	I do something to feel adrenaline
es_frequency_89	89	I act with hostility toward others
es_frequency_90	90	I snap or raise my voice at others when irritated
es_frequency_91	91	I show anger or irritability in response to minor slights or insults
es_frequency_92	92	I get back at someone who I feel has wronged me
es_frequency_96	96	I talk a lot
es_frequency_100	100	I do things that put others in danger
es_frequency_101	101	I lose valuable goods or money because I decide things too quickly
es_frequency_102	102	I skip school, work, or meetings to satisfy sudden urges
es_frequency_112	112	I leave dirty clothes on the floor for 30 minutes or longer
es_frequency_113	113	I damage a relationship with a friend because of something irresponsible I do

es_frequency_114	114	I miss a rent or mortgage payment
es_frequency_115	115	I do things that may cause me trouble with the law
es_frequency_123	123	I get grumpy
es_frequency_124	124	I feel like I might snap
es_frequency_125	125	Other people get on my nerves
es_frequency_126	126	Things bother me
es_frequency_127	127	I feel irritable
es_frequency_128	128	I lose my temper
es_frequency_129	129	I throw or kick things out of anger or frustration
es_frequency_130	130	I have sudden mood changes
es_frequency_131	131	I overreact to things
es_frequency_136	136	I have difficulty organizing tasks or activities
es_frequency_142	142	I perform poorly at school or work
es_frequency_150	150	I pretend to be angrier than I really am about someone's behavior in order to get them to behave differently in the future
es_frequency_151	151	I use emotional skills to make others feel guilty
es_frequency_152	152	I make someone feel ashamed about something that they have done in order to stop them from doing it again
es_frequency_153	153	I pay someone compliments so that they are more likely to do things for me in the future
es_frequency_168	168	I argue
es_frequency_169	169	I deliberately annoy others
es_frequency_170	170	I get in power struggles with others
es_frequency_175	175	I physically fight someone
es_frequency_176	176	I hit, kick, or push someone
es_frequency_177	177	I throw something at someone who angers me
es_frequency_178	178	I carry a weapon to use in a fight
es_frequency_187	187	I break rules
es_frequency_188	188	I cause problems by not following rules
es_frequency_189	189	I perform acts that are grounds for arrest
es_frequency_192	192	I gossip

es_frequency_193	193	I pit people against each other
es_frequency_194	194	I avoid talking to someone because I am upset with them
es_frequency_195	195	I say mean things (that may or may not be true) about others when they are not around
es_frequency_196	196	I isolate or ostracize others from a group
es_frequency_197	197	I post, re-post, or text an embarrassing photo or video of someone that they do not want others to see
es_frequency_198	198	I intentionally ignore someone's phone calls or text messages to hurt their feelings
es_frequency_200	200	I send a sexually suggestive message or picture to another that they do not want to receive
es_frequency_201	201	I pressure someone to send sexual or naked photos of themselves to me
es_frequency_203	203	I threaten to tell others a secret or lie to convince someone to have sex
es_frequency_205	205	I threaten to break up with someone if they don't have sex
es_frequency_214	214	I drink four or more (five or more for men) alcoholic beverages on one occasion?
es_frequency_215	215	I vape
es_frequency_216	216	I use cannabis/pot/marijuana
es_frequency_217	217	I use tobacco
es_frequency_218	218	I use illegal drugs
es_frequency_219	219	I use medicine without a doctor's prescription to change the way I feel
es_frequency_220	220	I drive while intoxicated
es_frequency_221	221	I do risky things to get drugs
es_frequency_222	222	I sell drugs to others
es_frequency_223	223	On how many occasions do you drink any kind of alcoholic beverage (e.g., beer, wine, liquor, etc.)?
es_frequency_224	224	On these occasions, how many alcoholic beverages did you drink per occasion on average?
es_frequency_226	226	I take an item from a store without paying for it
es_frequency_227	227	I download files illegally
es_frequency_229	229	I take items that are more expensive than what I paid for
es_frequency_230	230	I take things when others aren't looking
es_frequency_231	231	I buy things using a stolen credit card or credit card number
es_frequency_232	232	I embezzle money
es_frequency_233	233	I break into someone else's home, car, or building
es_frequency_239	239	I get angry when others provoke me

es_frequency_240	240	I get into verbal fights
es_frequency_241	241	I shout or make loud noises angrily
es_frequency_242	242	I curse or swear at people
es_frequency_243	243	I insult others
es_frequency_244	244	I tease others
es_frequency_245	245	I verbally assault others
es_frequency_246	246	I make hurtful comments to others
es_frequency_247	247	I make hurtful comments about others
es_frequencyLifetime_35	35	I defrauded others for money
es_frequencyLifetime_55	55	I deliberately started fires that caused damage
es_frequencyLifetime_103	103	I did something on impulse that led to others getting badly hurt or killed
es_frequencyLifetime_117	117	I did not show up to court when I was supposed to
es_frequencyLifetime_118	118	I was fired from a job
es_frequencyLifetime_179	179	I used a weapon against someone
es_frequencyLifetime_206	206	I forced someone into sexual activity
es_frequencyLifetime_207	207	I slipped someone drugs so that I could take advantage of them
es_frequencyLifetime_211	211	I contracted a sexually transmitted infection
es_frequencyLifetime_212	212	I engaged in sexual activity that resulted in an unplanned pregnancy
es_frequencyLifetime_225	225	I overdosed on drugs
es_frequencyLifetime_234	234	I stole while confronting a victim

Note. For item 224, we computed how many drinks a person drank (per unit time) by multiplying (a) the number of occasions in which they drank (per unit time; i.e., their response to item 223) and (b) the number of drinks they had per occasion (i.e., their response to item 224). Items above the dashed line reflect number of times engaging in the behavior per year. Items below the line reflect number of times engaging in the behavior over the lifetime. Many of the items were adapted/reprinted with permission from "Efficiently measuring dimensions of the externalizing spectrum model: Development of the externalizing spectrum inventory-computerized adaptive test (ESI-CAT)" by M. Sunderland, T. Slade, R. F. Krueger, K. E. Markon, C. J. Patrick, M. D. Kramer, 2017, *Psychological Assessment*, *29*(7), Supplementary Table 1 (https://doi.org/10.1037/pas0000384). Copyright 2017 by American Psychological Association.

Supplementary Table S6.

Descriptive Statistics of Externalizing Problem Items

Item	*n*	Missingness (%)	*M*	*SD*	Proportion Zero	Skewness (log)	Kurtosis (log)	Change
1	1237	0.00	3.23	9.45	.359	0.31	−1.31	−
2	1237	0.00	2.68	9.13	.368	0.46	−1.04	−
6	1237	0.00	20.89	141.68	.361	0.07	−1.45	−
7	1237	0.00	1.01	4.95	.600	1.31	0.48	−
8	1237	0.00	0.52	2.21	.725	1.90	2.48	−
10	1236	0.08	48.31	69.40	.081	−1.09	0.73	−
14	1237	0.00	0.00	0.12	.994	21.34	517.65	
15	1237	0.00	0.09	1.13	.957	6.33	42.06	−
16	1237	0.00	0.20	1.10	.824	2.77	6.87	−
17	1237	0.00	0.12	1.26	.941	5.25	28.04	−
18	1237	0.00	0.06	0.95	.983	9.51	94.84	−
27	1237	0.00	7.80	26.48	.231	−0.17	−1.14	−
28	1237	0.00	0.40	1.62	.665	1.60	1.29	−
29	1237	0.00	0.36	1.61	.779	2.11	3.30	−
30	1237	0.00	2.07	6.20	.317	0.16	−1.30	−
31	1237	0.00	0.39	2.88	.856	2.92	7.93	−
32	1237	0.00	0.06	1.01	.969	7.67	62.12	
33	1237	0.00	0.31	2.80	.826	2.62	5.82	−
34	1237	0.00	4.33	16.60	.366	0.29	−1.28	−
39	1237	0.00	1.33	4.33	.315	0.50	−0.98	−
40	1237	0.00	1.52	10.26	.662	1.49	1.07	−
41	1237	0.00	3.46	15.97	.408	0.68	−0.70	−
42	1237	0.00	0.44	3.27	.847	3.40	11.50	
51	1237	0.00	0.01	0.09	.987	15.25	263.59	
52	1237	0.00	0.02	0.49	.985	10.60	118.43	
53	1237	0.00	0.01	0.16	.996	20.45	436.04	
54	1237	0.00	0.02	0.22	.982	10.91	126.41	
58	1237	0.00	1.46	12.94	.793	2.42	5.10	−
59	1237	0.00	3.26	6.65	.448	0.33	−1.51	−
60	1237	0.00	5.58	46.22	.674	1.41	0.62	∩
61	1237	0.00	0.11	1.37	.959	6.60	46.38	∩
62	1237	0.00	0.03	0.31	.945	6.40	46.18	
64	1237	0.00	42.85	92.28	.099	−0.82	0.09	∩
65	1237	0.00	22.14	73.09	.084	−0.17	−0.61	∩
70	1237	0.00	1.38	25.98	.805	2.49	5.43	−
71	1237	0.00	0.34	5.45	.946	6.05	39.12	

72	1237	0.00	1.01	9.86	.876	3.45	11.55	−
76	1237	0.00	0.62	5.00	.731	2.19	3.87	−
77	1237	0.00	3.06	14.69	.484	0.86	−0.59	−
78	1237	0.00	1.32	11.47	.669	1.87	2.54	−
79	1237	0.00	1.94	14.78	.745	1.91	2.34	−
80	1237	0.00	0.51	5.18	.827	2.99	8.60	−
81	1237	0.00	2.66	13.53	.556	1.11	−0.07	−
89	1237	0.00	1.15	5.39	.604	1.28	0.45	−
90	1237	0.00	3.63	21.88	.288	0.14	−1.26	∩
91	1237	0.00	2.59	10.40	.455	0.61	−0.95	∩
92	1237	0.00	0.33	1.71	.768	2.25	4.01	−
96	1237	0.00	123.62	1020.45	.123	−0.55	−0.49	−
100	1237	0.00	0.25	2.56	.893	4.21	18.25	
101	1237	0.00	1.54	11.71	.578	1.35	0.67	−
102	1237	0.00	1.45	4.98	.648	1.29	0.18	−
112	1237	0.00	14.12	32.96	.308	−0.19	−1.47	−
113	1237	0.00	0.09	0.59	.845	3.76	15.46	−
114	1237	0.00	0.02	0.15	.953	6.04	41.24	
115	1237	0.00	1.43	15.96	.825	2.74	6.60	−
123	1237	0.00	16.09	89.24	.061	−0.59	0.31	−
124	1237	0.00	6.39	23.60	.289	0.22	−1.13	∩
125	1237	0.00	22.98	54.83	.046	−0.54	0.10	−
126	1237	0.00	43.63	149.09	.037	−0.67	0.62	∩
127	1237	0.00	25.47	100.49	.053	−0.63	0.39	∩
128	1237	0.00	2.75	13.56	.345	0.56	−0.74	
129	1237	0.00	0.58	4.77	.734	2.41	5.25	−
130	1237	0.00	9.16	31.70	.375	0.32	−1.18	−
131	1237	0.00	8.71	32.65	.231	0.12	−0.95	−
136	1237	0.00	17.90	54.42	.227	−0.08	−1.10	∩
142	1237	0.00	4.35	16.70	.419	0.52	−1.08	∩
150	1237	0.00	1.72	25.80	.651	1.27	0.25	−
151	1237	0.00	0.99	8.89	.669	1.44	0.75	−
152	1237	0.00	1.16	9.53	.608	1.26	0.26	∩
153	1237	0.00	2.98	29.45	.643	1.35	0.49	∩
168	1237	0.00	11.13	87.05	.093	−0.54	−0.25	+
169	1237	0.00	8.69	171.26	.559	0.91	−0.52	−
170	1237	0.00	2.58	14.56	.507	0.90	−0.34	
175	1237	0.00	0.02	0.51	.976	10.32	121.94	
176	1237	0.00	0.32	5.57	.922	5.17	28.49	−
177	1237	0.00	0.06	0.91	.956	6.82	49.03	−
178	1237	0.00	0.30	2.91	.982	8.76	76.88	

187	1237	0.00	4.27	19.77	.537	0.88	−0.63	−
188	1237	0.00	0.50	2.93	.814	2.48	5.24	−
189	1237	0.00	1.13	14.13	.886	3.66	12.99	−
192	1237	0.00	19.71	98.47	.169	−0.25	−0.81	−
193	1237	0.00	0.24	3.01	.899	4.29	18.98	−
194	1237	0.00	3.45	13.32	.232	0.04	−1.10	−
195	1237	0.00	3.30	19.35	.533	0.86	−0.62	−
196	1237	0.00	0.22	1.62	.886	3.76	13.93	−
197	1237	0.00	0.21	4.32	.929	5.13	27.42	−
198	1237	0.00	0.83	6.52	.812	2.65	6.45	
200	1237	0.00	0.00	0.03	.994	18.18	383.91	
201	1237	0.00	0.00	0.03	.998	29.58	919.21	
205	1237	0.00	0.00	0.03	.998	29.94	958.07	
214	1237	0.00	1.50	4.17	.569	0.95	−0.61	−
215	1237	0.00	33.49	186.49	.804	2.30	4.03	−
216	1237	0.00	4.04	16.86	.682	1.64	1.39	∩
217	1237	0.00	12.93	81.33	.884	3.58	12.09	∩
218	1237	0.00	2.49	26.24	.935	6.03	37.98	∩
219	1237	0.00	0.67	17.20	.957	6.82	50.25	
220	1237	0.00	0.06	0.49	.924	4.88	25.32	
221	1237	0.00	0.28	8.54	.981	8.79	79.22	
222	1237	0.00	0.70	21.40	.981	9.73	101.98	∩
223	1237	0.00	5.99	10.54	.277	−0.20	−1.39	+
224	1237	0.00	18.22	37.83	.277	−0.18	−1.41	+
226	1237	0.00	0.09	0.61	.909	4.66	23.01	−
227	1237	0.00	0.51	3.98	.878	3.51	12.21	
229	1237	0.00	0.05	0.44	.951	6.07	38.83	−
230	1237	0.00	0.05	0.51	.953	6.82	50.29	
231	1237	0.00	0.04	1.02	.998	24.57	606.63	
233	1237	0.00	0.00	0.01	.997	18.87	373.58	
239	1237	0.00	2.41	9.02	.395	0.58	−0.82	−
240	1237	0.00	1.11	4.46	.562	1.15	0.15	
241	1237	0.00	1.92	11.08	.563	1.21	0.25	
242	1237	0.00	18.85	120.10	.517	1.06	−0.10	−
243	1237	0.00	4.26	86.07	.682	1.74	2.08	−
244	1237	0.00	23.00	134.11	.324	0.09	−1.28	−
245	1237	0.00	0.31	2.67	.922	4.53	20.83	
246	1237	0.00	0.63	4.30	.810	2.55	5.72	∩
247	1237	0.00	1.18	5.55	.708	1.63	1.28	−
35	1237	0.00	0.72	14.82	.977	10.73	132.90	
55	1237	0.00	0.01	0.11	.993	12.49	162.74	+

103	1237	0.00	0.02	0.24	.986	10.47	125.74	
117	1237	0.00	0.06	0.50	.973	7.39	59.43	∩
118	1237	0.00	0.38	1.34	.804	2.51	7.50	+
179	1237	0.00	0.09	1.60	.982	9.20	91.84	+
206	1237	0.00	0.01	0.16	.994	13.51	188.97	+
211	1237	0.00	0.28	0.98	.836	2.91	10.21	+
212	1237	0.00	0.35	2.92	.833	2.32	4.54	+
225	1237	0.00	0.10	0.71	.963	5.98	37.55	
234	1237	0.00	0.02	0.40	.998	23.58	566.01	

Note. The descriptive statistics in the table are based on the rate of engaging the behavior per month, whereas the items were modeled in terms of their frequency rate per year. The "Proportion Zero" column indicates the proportion of scores for that item that indicated a frequency count of zero. The "Change" column indicates significant age-related differences as a function of age from a generalized additive model: "+"indicates increases as a function of age (i.e., peak in older adulthood); "−" indicates decreases as a function of age (i.e., peak in early adulthood); "∩" indicates an inverted-U-shaped pattern (i.e., increases and then decreases) as a function of age (i.e., peak in middle adulthood); no symbol indicates no significant age-related differences. Items above the dashed line reflect number of times engaging in the behavior per year. Items below the line reflect number of times engaging in the behavior over the lifetime.

Supplementary Table S7.

Percentiles for Behavior Frequencies by Externalizing Problem Item Per Year

	Percentile																				
Item	0	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
1	0	0	0	0	0	0	0	0	1	2	13	13	13	13	25	37	53	53	105	157	262
2	0	0	0	0	0	0	0	0	1	2	3	13	13	13	13	25	37	53	53	122	367
6	0	0	0	0	0	0	0	0	5	21	25	49	53	74	122	157	209	366	366	1096	2193
7	0	0	0	0	0	0	0	0	0	0	0	0	1	1	2	4	13	13	25	53	210
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	6	13	25	158
10	0	0	13	25	53	61	105	122	157	209	261	366	366	522	731	731	1096	1096	1461	1827	5480
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	53
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	62
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3	13	62
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	62
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	106
27	0	0	0	0	0	1	4	13	13	13	25	25	37	53	53	61	105	108	183	366	523
28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	2	13	13	25	62
29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3	13	25	158
30	0	0	0	0	0	0	0	1	2	4	6	13	13	13	25	25	37	49	53	105	262
31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	25	367
32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	54
33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3	13	62
34	0	0	0	0	0	0	0	0	2	3	10	13	13	25	25	37	53	53	105	209	315
39	0	0	0	0	0	0	0	1	1	2	2	3	5	13	13	13	25	25	37	53	208
40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3	13	13	37	61	732
41	0	0	0	0	0	0	0	0	0	1	2	3	6	13	13	25	25	37	53	108	367
42	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	13	367

51	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	25
52	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14
53	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	61
54	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	53
58	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	4	13	53	732
59	0	0	0	0	0	0	0	0	0	1	3	13	13	25	37	53	53	105	105	157	1000
60	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	13	25	53	105	366	1828
61	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	101
62	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	62
64	0	0	1	13	25	53	53	105	105	157	157	209	261	366	366	731	731	1096	1461	1827	5480
65	0	0	1	2	13	13	13	25	25	49	53	53	105	105	105	157	261	366	731	1096	2558
70	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	13	25	367
71	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	210
72	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	25	1097
76	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	3	13	25	184
77	0	0	0	0	0	0	0	0	0	0	1	2	3	5	13	13	25	37	61	157	306
78	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	2	10	13	39	160
79	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	13	25	61	262
80	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3	13	158
81	0	0	0	0	0	0	0	0	0	0	0	0	1	3	5	13	25	37	53	122	367
89	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	10	13	13	25	53	306
90	0	0	0	0	0	0	1	2	3	5	13	13	13	13	25	25	37	53	61	105	210
91	0	0	0	0	0	0	0	0	0	0	2	3	13	13	13	25	25	49	53	105	367
92	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	13	13	75
96	0	0	0	4	13	25	53	53	105	105	157	209	261	366	366	731	1096	1096	1827	3653	7306
100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3	106
101	0	0	0	0	0	0	0	0	0	0	0	0	1	1	2	4	13	13	25	53	210
102	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	5	13	25	53	105	262
112	0	0	0	0	0	0	0	7	13	25	53	53	105	105	157	244	366	366	366	731	2923

113	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	2	123
114	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	53
115	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	13	53	732
123	0	0	4	13	13	25	25	37	49	53	53	74	105	105	157	157	168	261	366	731	1462
124	0	0	0	0	0	0	1	2	3	13	13	13	13	25	37	50	53	105	122	314	549
125	0	1	5	13	25	25	37	49	53	61	105	105	110	157	183	261	366	366	731	1096	1462
126	0	3	13	25	49	53	53	74	105	105	157	157	209	261	366	366	731	731	1096	1827	2923
127	0	0	13	13	25	25	37	53	53	61	105	105	105	157	186	261	366	366	731	1096	1567
128	0	0	0	0	0	0	0	1	1	2	2	5	13	13	13	13	25	37	53	105	419
129	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	2	13	25	315
130	0	0	0	0	0	0	0	0	2	5	13	13	25	25	49	53	105	157	209	366	1097
131	0	0	0	0	0	1	2	4	13	13	13	25	25	37	49	53	84	105	209	366	1097
136	0	0	0	0	0	1	6	13	13	25	37	53	53	98	105	157	209	366	522	1096	1828
142	0	0	0	0	0	0	0	0	0	1	2	5	13	13	25	25	37	53	105	157	367
150	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	13	13	25	25	53	367
151	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3	13	13	25	50	184
152	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	4	13	13	25	53	184
153	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	6	13	25	53	105	315
168	0	0	1	3	13	13	13	25	25	25	37	53	53	53	74	105	122	157	261	366	523
169	0	0	0	0	0	0	0	0	0	0	0	0	2	13	13	25	37	53	105	209	732
170	0	0	0	0	0	0	0	0	0	0	0	1	3	13	13	13	25	37	53	105	732
175	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	54
176	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	210
177	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	38
178	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	366
187	0	0	0	0	0	0	0	0	0	0	0	1	2	10	13	25	37	57	105	261	732
188	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	13	25	210
189	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	25	836
192	0	0	0	0	2	6	13	25	25	37	53	53	61	105	105	157	209	366	366	1054	2558

193	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3	106
194	0	0	0	0	0	1	2	3	5	10	13	13	13	25	25	37	49	53	61	151	262
195	0	0	0	0	0	0	0	0	0	0	0	1	3	12	13	25	37	53	105	157	732
196	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	13	106
197	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	106
198	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	13	25	367
200	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13
201	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13
205	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13
214	0	0	0	0	0	0	0	0	0	0	0	0	1	3	10	13	25	37	53	105	367
215	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13	122	1827	36525
216	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3	13	37	105	314	2193
217	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	105	18263
218	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	7305
219	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	367
220	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	62
221	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	62
222	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	210
223	0	0	0	0	0	0	2	5	13	25	25	37	49	53	61	105	105	157	209	366	732
224	0	0	0	0	0	0	2	11	18	30	49	74	105	147	185	244	314	418	627	1044	2610
226	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	105
227	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	13	367
229	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	54
230	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	105
231	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	366
233	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
239	0	0	0	0	0	0	0	0	1	2	2	4	13	13	13	25	25	48	53	105	523
240	0	0	0	0	0	0	0	0	0	0	0	0	1	2	3	13	13	25	25	53	367
241	0	0	0	0	0	0	0	0	0	0	0	0	1	2	3	13	13	25	49	105	367

242	0	0	0	0	0	0	0	0	0	0	0	1	3	13	13	25	53	105	261	1096	2193
243	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	13	13	25	64	732
244	0	0	0	0	0	0	0	2	13	25	25	37	53	61	105	122	209	366	501	1096	2558
245	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	262
246	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	13	25	367
247	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	5	13	37	53	210
35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	500
55	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
103	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6
117	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10
118	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	2	30
179	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6
206	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
211	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	20
212	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	2	6
225	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6
234	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10

Note. Items above the dashed line reflect number of times engaging in the behavior per year. Items below the line reflect number of times engaging in the behavior over the lifetime.

Supplementary Table S8.

Identifying the Best-Fitting Form of Age-Related Differences: Estimates of Model Fit from the Bayesian Item Response Theory Models

Model	ELPD Difference	*SE* of ELPD Difference	ELPD	*SE* of ELPD	Effective Number of Parameters	*SE* of Effective Number of Parameters	WAIC	*SE* of WAIC
Generalized Additive Model	0.00	0.00	−311725.31	1090.09	2052.37	45.50	623450.62	2180.18
Linear	−0.68	1.83	−311725.99	1090.10	2052.44	45.54	623451.98	2180.19
Quadratic	−2.24	1.60	−311727.55	1090.07	2053.88	45.42	623455.10	2180.14
Null Model (No Predictors)	−2.89	6.69	−311728.20	1089.86	2056.35	45.12	623456.40	2179.73

Note. “ELPD” = expected log predictive density (smaller is better), calculated based on WAIC; “WAIC” widely applicable information criterion (smaller is better); “ELPD Difference” = difference of model fit (ELPD) relative to generalized additive model; “SE” = standard error.

Supplementary Table S9.

Item Characteristics From the Bayesian Item Response Theory Model

Item	Easiness (β)	Discrimination (α)	mu (μ)	phi (φ)	ZI
1	5.11	0.88	41.00	0.41	0.18
2	5.17	1.05	33.22	0.35	0.12
6	6.55	0.59	271.89	0.51	0.32
7	4.72	1.60	8.95	0.21	0.08
8	4.76	1.98	5.15	0.16	0.15
10	7.56	0.64	695.67	0.81	0.07
14	2.20	1.99	0.90	0.06	0.81
15	3.32	3.29	0.18	0.07	0.26
16	4.17	2.39	1.50	0.12	0.10
17	3.11	2.12	0.93	0.06	0.33
18	3.39	2.97	0.47	0.07	0.56
27	5.64	0.72	89.83	0.63	0.17
28	3.52	1.32	4.15	0.17	0.04
29	4.41	1.80	4.84	0.10	0.12
30	4.77	0.88	29.08	0.71	0.20
31	4.37	1.86	4.29	0.06	0.16
32	1.54	0.66	2.28	0.04	0.54
33	3.20	1.37	2.85	0.08	0.11
34	5.70	1.17	47.05	0.55	0.23
39	4.09	0.84	15.72	0.38	0.02
40	5.63	1.92	13.54	0.18	0.18
41	4.87	0.94	29.24	0.22	0.02
42	4.29	1.84	4.02	0.05	0.06
51	1.03	2.49	0.08	0.07	0.48
52	2.23	2.43	0.31	0.06	0.55
53	3.34	2.56	1.92	0.05	0.83
54	2.73	2.95	0.21	0.04	0.41
58	6.14	2.76	5.99	0.09	0.08
59	5.39	0.75	66.89	0.73	0.40
60	6.37	1.29	76.24	0.16	0.39
61	3.35	3.25	0.19	0.05	0.23
62	1.22	1.42	0.39	0.04	0.20
64	7.76	0.88	575.81	0.70	0.07
65	6.79	0.80	248.57	0.41	0.00
70	5.25	2.50	3.74	0.10	0.10
71	4.59	3.08	0.84	0.03	0.19

72	5.34	1.79	12.72	0.03	0.12
76	4.54	2.01	3.94	0.13	0.04
77	5.42	1.38	25.39	0.21	0.05
78	3.83	1.19	7.00	0.12	0.03
79	6.27	2.61	8.74	0.10	0.05
80	4.64	2.59	1.76	0.10	0.06
81	5.47	1.43	24.57	0.16	0.05
89	5.22	1.77	11.19	0.34	0.26
90	4.82	0.84	32.53	0.64	0.17
91	5.47	1.28	30.96	0.60	0.31
92	3.87	1.81	2.74	0.12	0.06
96	7.56	0.54	807.16	0.40	0.06
100	3.51	2.10	1.24	0.05	0.11
101	4.29	1.24	10.25	0.19	0.07
102	5.99	1.88	20.51	0.29	0.36
112	6.15	0.40	246.74	0.64	0.29
113	2.32	1.49	0.96	0.09	0.04
114	0.99	0.98	0.73	0.10	0.47
115	5.23	1.97	8.34	0.06	0.11
123	6.62	0.95	166.44	1.04	0.04
124	6.11	1.38	50.38	0.49	0.09
125	7.07	0.99	243.27	0.84	0.01
126	7.51	0.87	457.81	0.79	0.01
127	7.00	0.94	246.21	0.92	0.02
128	5.17	1.27	23.62	0.35	0.02
129	4.43	1.85	4.50	0.11	0.03
130	7.01	1.56	93.27	0.52	0.24
131	6.50	1.37	75.60	0.45	0.04
136	6.84	0.91	222.51	0.35	0.09
142	5.56	1.13	43.25	0.31	0.16
150	4.39	0.84	21.26	0.29	0.45
151	4.63	1.45	10.33	0.19	0.22
152	4.23	1.21	10.08	0.16	0.04
153	4.64	0.85	27.15	0.15	0.25
168	5.66	0.65	103.11	0.75	0.04
169	6.61	1.70	50.04	0.35	0.36
170	5.23	1.19	28.52	0.24	0.16
175	1.88	2.92	0.08	0.05	0.28
176	3.96	2.42	1.19	0.04	0.10
177	2.87	2.82	0.23	0.05	0.21
178	5.28	1.22	54.35	0.04	0.69

187	5.28	0.72	62.69	0.23	0.30
188	4.44	1.86	4.62	0.11	0.22
189	6.26	3.18	3.56	0.04	0.09
192	7.06	1.08	209.37	0.56	0.10
193	3.82	2.39	1.08	0.06	0.13
194	4.92	0.86	34.86	0.52	0.06
195	6.15	1.51	42.79	0.35	0.33
196	4.12	2.24	1.83	0.05	0.11
197	4.27	2.86	0.82	0.04	0.14
198	4.10	1.41	6.48	0.06	0.08
200	1.56	1.71	0.66	0.10	0.82
201	2.27	2.39	2.50	0.07	0.88
205	1.03	2.27	0.60	0.15	0.82
214	4.37	0.53	33.90	0.36	0.44
215	7.67	0.66	809.05	0.05	0.39
216	4.94	0.55	58.92	0.07	0.09
217	5.66	0.07	295.08	0.02	0.30
218	6.37	2.74	8.58	0.02	0.17
219	4.09	1.78	4.44	0.02	0.35
220	0.85	0.43	1.33	0.04	0.27
221	3.29	3.52	0.15	0.04	0.36
222	3.69	4.37	0.06	0.05	0.32
223	4.55	−0.01	95.06	0.65	0.25
224	6.13	0.26	305.14	0.48	0.24
226	2.98	2.06	0.78	0.06	0.13
227	3.33	0.77	8.54	0.03	0.15
229	2.36	1.72	0.85	0.06	0.37
230	1.65	0.99	1.24	0.02	0.27
231	3.07	3.26	1.90	0.07	0.82
233	−0.71	2.48	0.03	0.25	0.58
239	5.40	1.35	25.81	0.41	0.13
240	4.76	1.45	11.66	0.23	0.09
241	4.90	1.35	15.90	0.18	0.06
242	7.60	1.90	98.69	0.13	0.02
243	6.20	2.48	9.84	0.15	0.08
244	7.26	1.06	263.10	0.46	0.25
245	4.73	2.94	1.16	0.05	0.17
246	4.73	2.05	4.53	0.08	0.09
247	5.17	1.61	13.95	0.19	0.32
35	3.22	2.48	0.69	0.02	0.38
55	−1.05	0.57	0.35	0.25	0.78

103	−0.86	1.14	0.11	0.17	0.53
117	−0.07	0.82	0.37	0.21	0.59
118	−0.69	0.08	0.45	0.23	0.08
179	−0.35	0.66	0.37	0.09	0.64
206	1.04	1.25	0.78	0.14	0.83
211	−1.23	−0.01	0.30	0.23	0.07
212	−0.42	0.13	0.57	1.08	0.41
225	−0.50	0.72	0.25	0.14	0.43
234	2.48	2.09	4.56	0.06	0.88

Note. “ZI” = zero-inflation probability. Items above the dashed line reflect number of times engaging in the behavior per year. Items below the line reflect number of times engaging in the behavior over the lifetime.

Supplementary Table S10.

Item–Construct Correlations

Item	EFA Loading	Item–Total Correlation	Item–Latent Correlation
1	.49	.53	.46
2	.50	.53	.45
6	.41	.40	.39
7	.58	.60	.46
8	.50	.52	.39
10	.42	.44	.51
14	.17	.12	.07
15	.42	.38	.23
16	.50	.50	.35
17	.41	.38	.22
18	.29	.26	.15
27	.46	.47	.45
28	.53	.54	.39
29	.43	.45	.33
30	.55	.58	.50
31	.42	.41	.28
32	.12	.08	.07
33	.42	.41	.28
34	.55	.58	.52
39	.50	.51	.45
40	.54	.55	.44
41	.44	.43	.41
42	.37	.35	.28
51	.24	.20	.11
52	.19	.16	.10
53	.14	.12	.06
54	.25	.22	.12
58	.61	.60	.40
59	.53	.55	.44
60	.53	.51	.38
61	.45	.40	.23
62	.36	.32	.18
64	.46	.48	.57
65	.44	.44	.48
70	.55	.56	.38
71	.34	.30	.20

72	.28	.27	.21
76	.52	.53	.38
77	.56	.58	.47
78	.51	.50	.36
79	.60	.60	.42
80	.51	.50	.36
81	.50	.51	.41
89	.59	.60	.48
90	.48	.48	.47
91	.57	.59	.52
92	.55	.55	.37
96	.42	.45	.46
100	.39	.37	.25
101	.52	.52	.39
102	.56	.58	.45
112	.34	.34	.35
113	.45	.43	.27
114	.25	.21	.14
115	.48	.45	.30
123	.52	.55	.63
124	.57	.60	.60
125	.51	.53	.66
126	.44	.46	.61
127	.50	.53	.64
128	.52	.53	.52
129	.47	.46	.35
130	.57	.60	.59
131	.58	.62	.61
136	.42	.42	.47
142	.45	.47	.46
150	.47	.49	.38
151	.51	.52	.41
152	.48	.49	.40
153	.39	.39	.32
168	.45	.46	.51
169	.51	.54	.47
170	.47	.48	.42
175	.36	.31	.16
176	.38	.36	.23
177	.35	.32	.19
178	.17	.14	.09

187	.52	.51	.40
188	.54	.53	.35
189	.47	.44	.30
192	.48	.52	.52
193	.39	.37	.25
194	.49	.51	.51
195	.51	.53	.47
196	.37	.37	.25
197	.28	.28	.20
198	.46	.44	.31
200	.13	.09	.05
201	.15	.11	.05
205	.17	.12	.06
214	.44	.43	.31
215	.33	.31	.23
216	.36	.33	.24
217	.21	.15	.09
218	.40	.32	.22
219	.24	.19	.12
220	.21	.17	.10
221	.36	.31	.17
222	.35	.29	.18
223	.23	.21	.13
224	.30	.28	.19
226	.39	.36	.24
227	.19	.17	.15
229	.31	.26	.17
230	.30	.25	.15
231	.27	.21	.08
233	.15	.13	.07
239	.58	.59	.54
240	.57	.57	.46
241	.52	.52	.43
242	.59	.59	.50
243	.64	.64	.48
244	.42	.44	.42
245	.47	.43	.27
246	.52	.50	.35
247	.55	.55	.39
35	.31	.23	.14
55	.00	−.04	−.02

103	.22	.15	.09
117	.23	.15	.10
118	.04	−.02	.01
179	.14	.08	.05
206	.02	−.01	.01
211	.03	−.03	−.07
212	.06	.00	.02
225	.24	.19	.14
234	.22	.14	.06

Note. “EFA” = exploratory factor analysis. Items above the dashed line reflect number of times engaging in the behavior per year. Items below the line reflect number of times engaging in the behavior over the lifetime.

Supplementary Table S11.

Validity Analyses by Age and Sex

Outcome: ASR Externalizing		
Group	*r*	Discriminant Validity (Fisher's *r*-to-*z*)
Men	.50***	$z = 2.79^{**}$
Women	.58***	$z = 3.57^{***}$
Young Adults	.52***	$z = 4.10^{***}$
Mid-Aged Adults	.57***	$z = 2.02^{*}$
Older Adults	.51***	$z = 1.59$

Outcome: Functional Impairment			
Group	*r*	β (controlling for ASR)	β (controlling for dichotomous items)
Men	.27***	0.05	0.12
Women	.37***	0.10**	0.39***
Young Adults	.35***	0.11**	0.36***
Mid-Aged Adults	.32***	0.06	0.21**
Older Adults	.43***	0.21***	0.27***

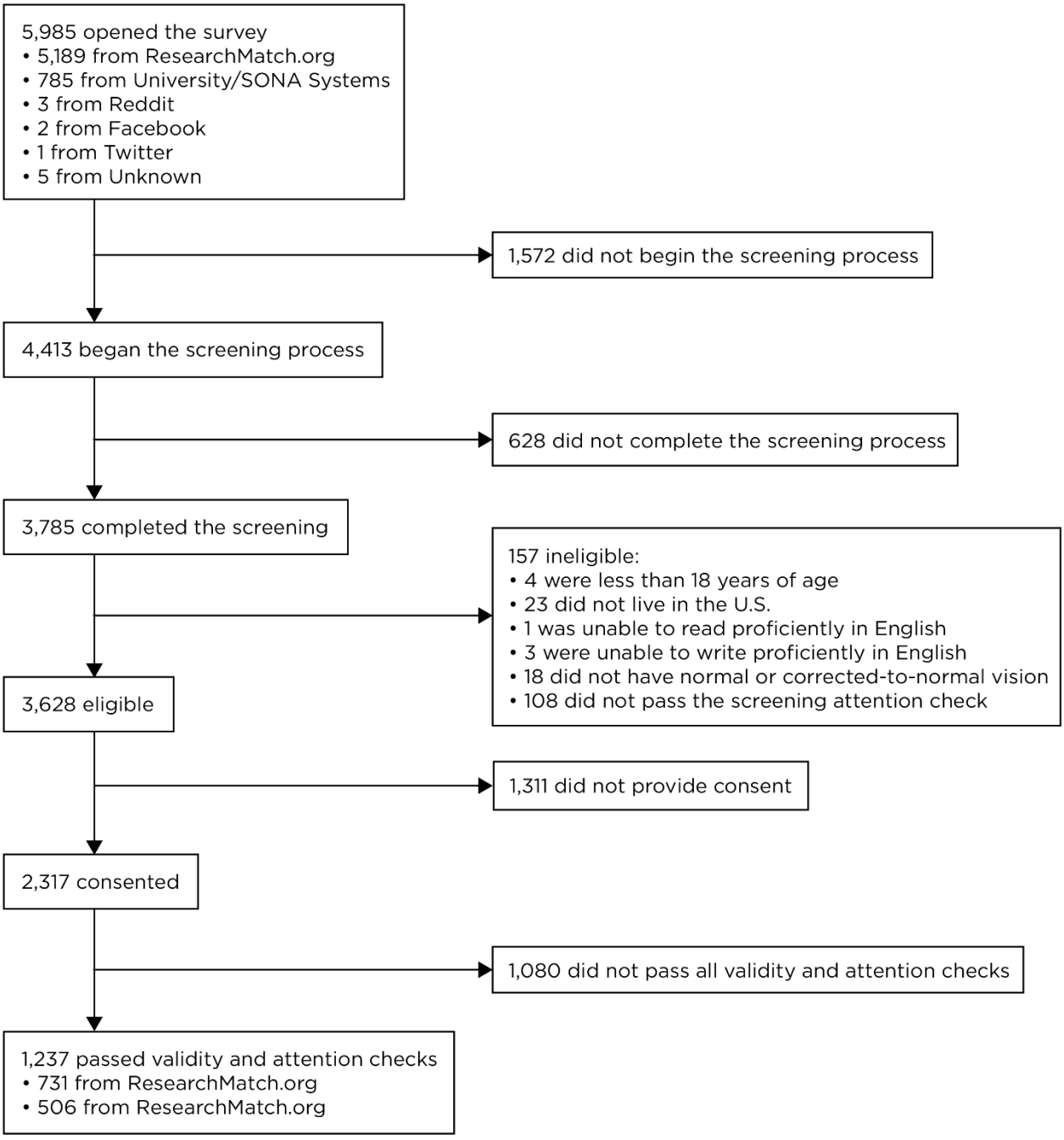
Outcome: Inhibitory Control			
Group	*r*	β (controlling for ASR)	β (controlling for dichotomous items)
Men	−.06	0.00	−0.06
Women	−.23***	−0.14*	−0.18**
Young Adults	−.10†	−0.09	−0.21**
Mid-Aged Adults	−.12	−0.10†	−0.10
Older Adults	−.07	0.00	−0.07

Note. The "*r*" column represents estimates of convergent and criterion-related validity of the latent externalizing problem scores from the Bayesian item response model in relation to the criterion (outcome variable), as operationalized with Pearson correlation. The "β" column represents estimates of incremental validity of the latent externalizing problem scores from the Bayesian item response model above and beyond the Adult Self-Report (ASR) or dichotomous versions of the same items in predicting the outcome. The "Discriminant Validity" column represents estimates of discriminant validity—i.e., whether the latent externalizing problem scores from the Bayesian item response model are more strongly associated with the ASR Externalizing scale than with the ASR Internalizing scale (based on a Fisher's *r*-to-*z* test). "dichotomous items" represents the latent externalizing problem scores from an item response theory model based on dichotomized items. Age groups were partitioned to retain at least 300 participants per group for considerations relating to power, while still separating meaningful developmental periods. "Young adults" refer to ages 18–25 ($n = 548$); "Mid-aged adults" refer to ages 25–55 ($n = 328$); "Older adults" refer to ages 55 and over ($n = 361$). The models predicting

inhibitory control include age as a covariate because of the known developmental changes in inhibitory control.

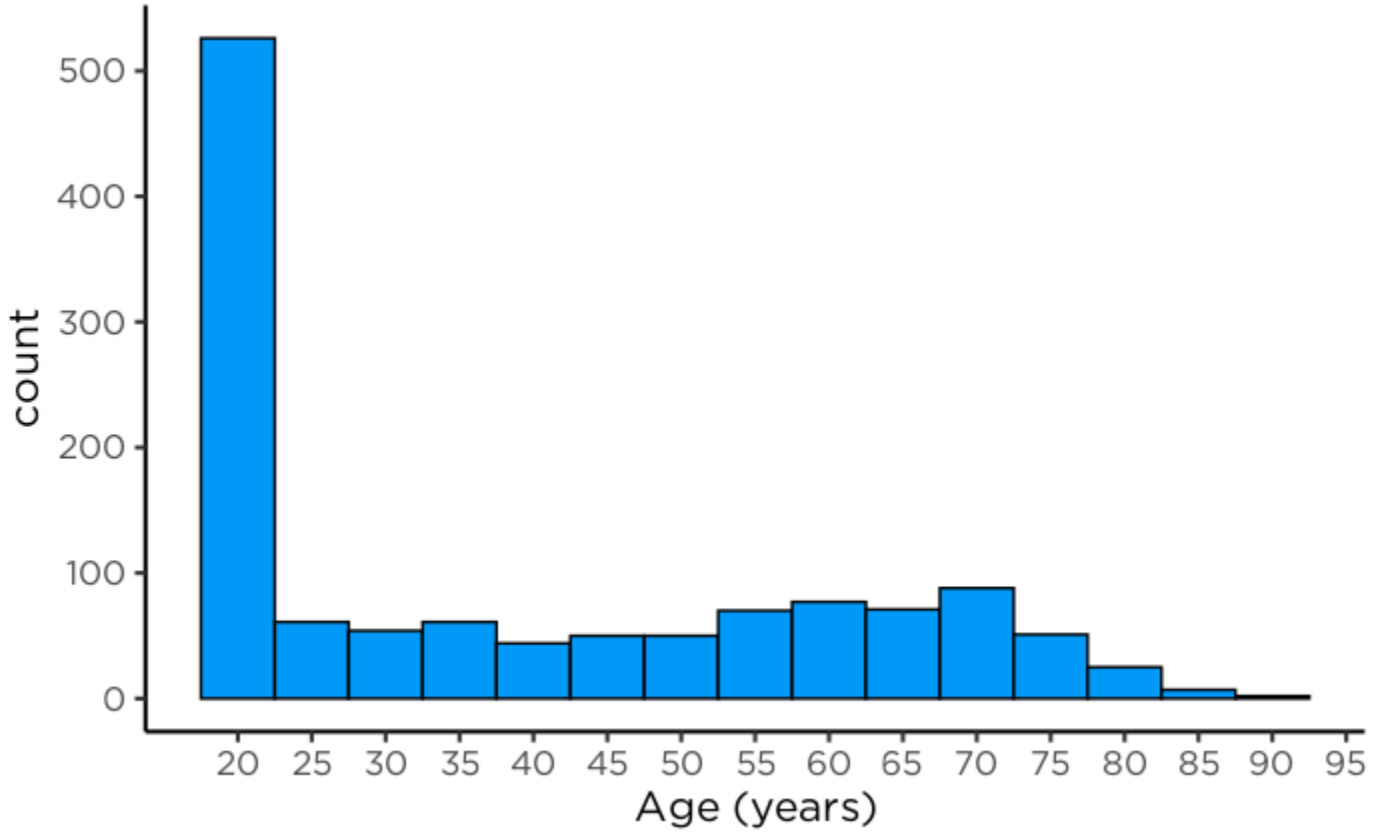
Supplementary Figure S1

Participant Flowchart



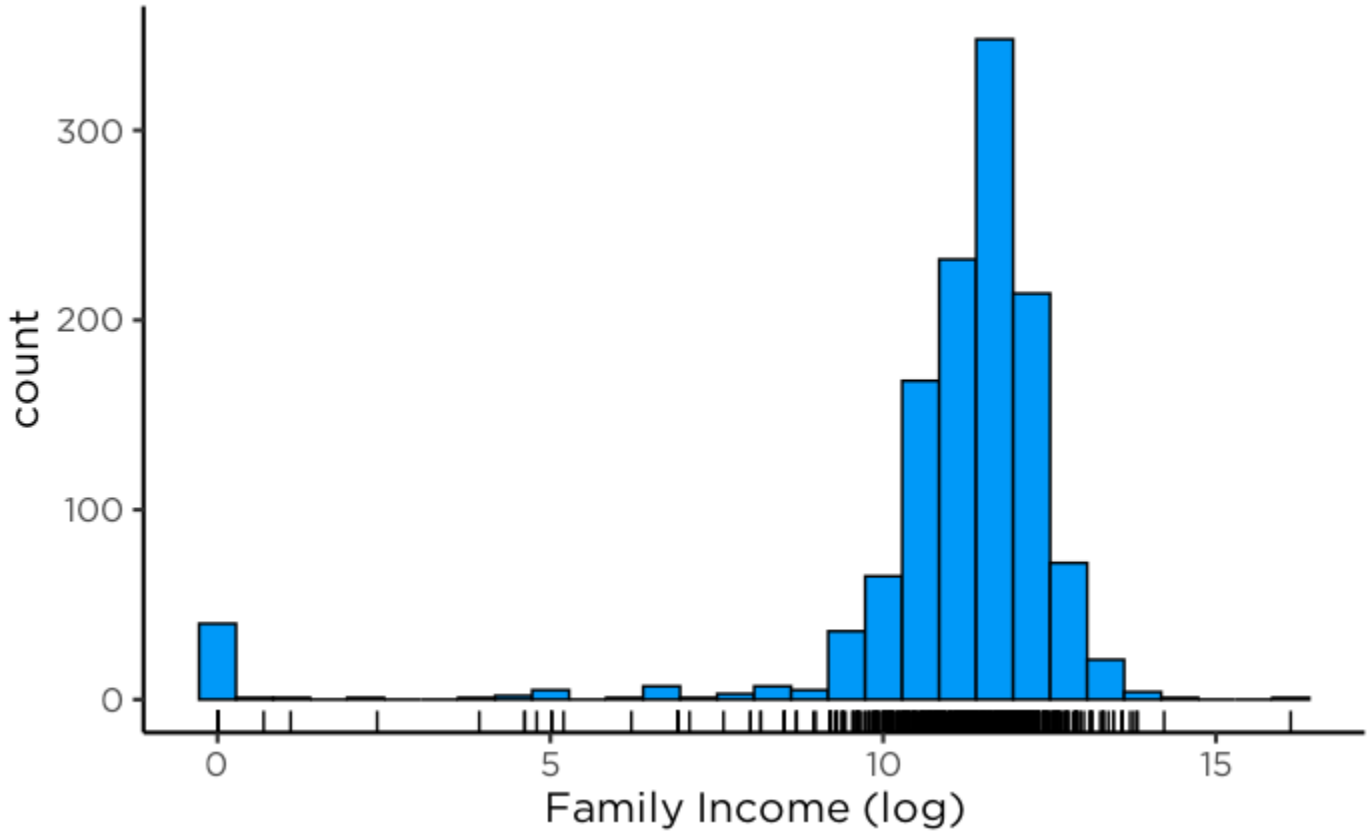
Supplementary Figure S2

Histogram of Participants' Age in the Sample



Supplementary Figure S3

Histogram of Participants' Family Income in the Sample



Note. A rug plot is depicted with black lines below the histogram and along the x-axis.

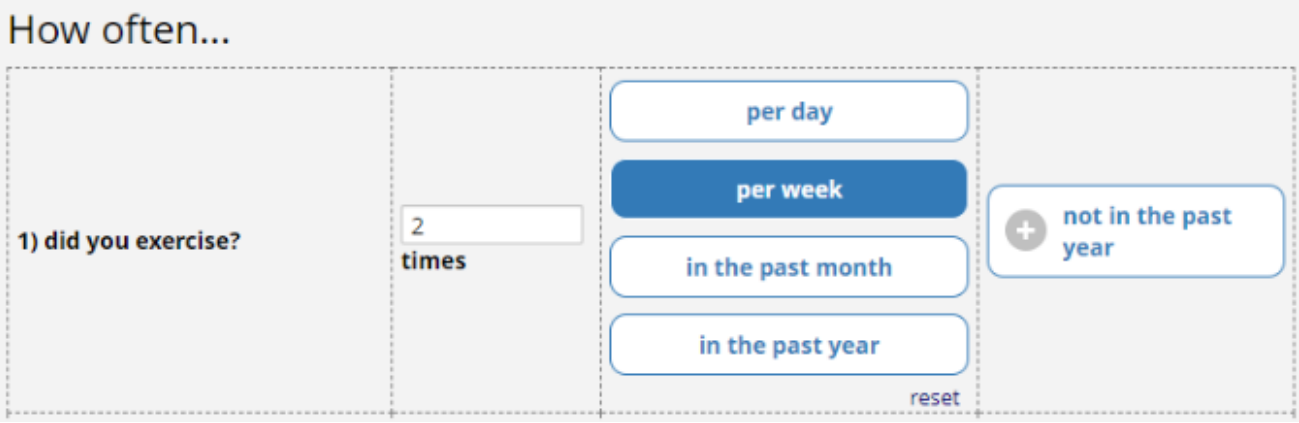
Supplementary Figure S4

Questionnaire Instructions

The following questions are meant to study a range of behaviors that many people engage in. When responding, please do your best to consider your actual behavior, not how you would ideally like to behave. Please answer all items as accurately as you can, even if some do not seem to apply to you. You can pause the survey and return later by pressing "save & return later" at the bottom of the screen.

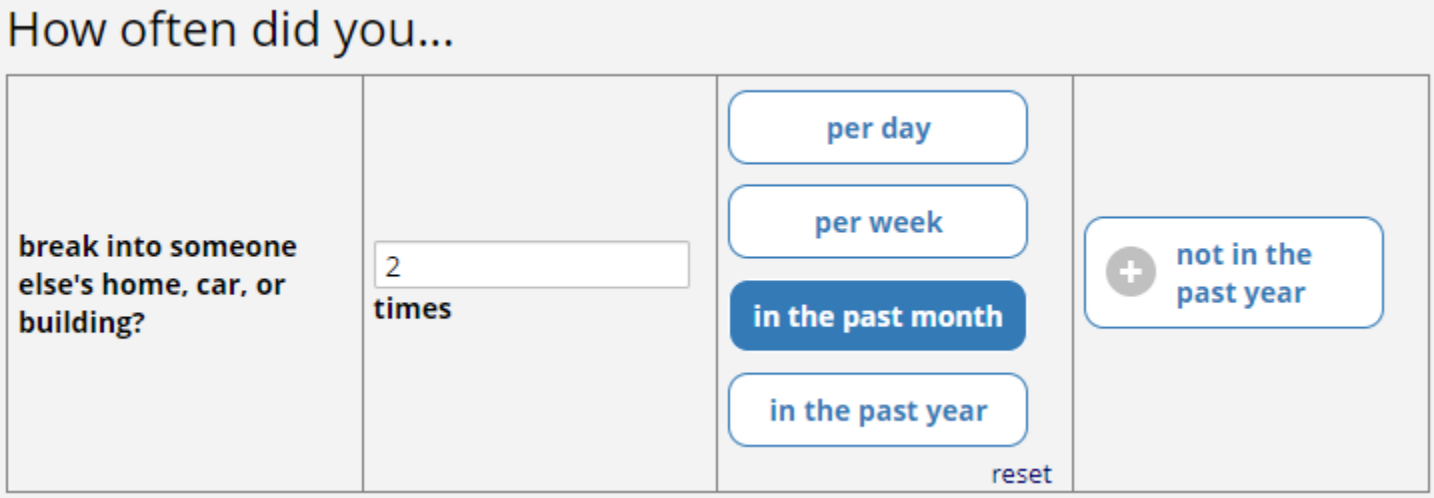
Please reflect on your behavior in the past month (30 days). Please select the most appropriate timeframe and indicate how many times you engaged in the listed behavior during that timeframe. If you select "*in the past month*", please indicate the total number of times you engaged in the behavior in the past month. It might make more sense to think about how often you engaged in a behavior on a daily or weekly basis; in this case, please indicate the average number of times you engaged in the behavior "*per day*" or "*per week*", using the past month (30 days) as your timeframe of reference. Finally, you might have engaged in a behavior in the past year but not in the past month; in this case, please indicate the total number of times you engaged in the behavior and select "*in the past year*". If you did not engage in the behavior in the past year, please select "*not in the past year*" instead of entering 0.

For example, if within the past month you exercised, on average, 2 times per week, you would indicate that as shown here:



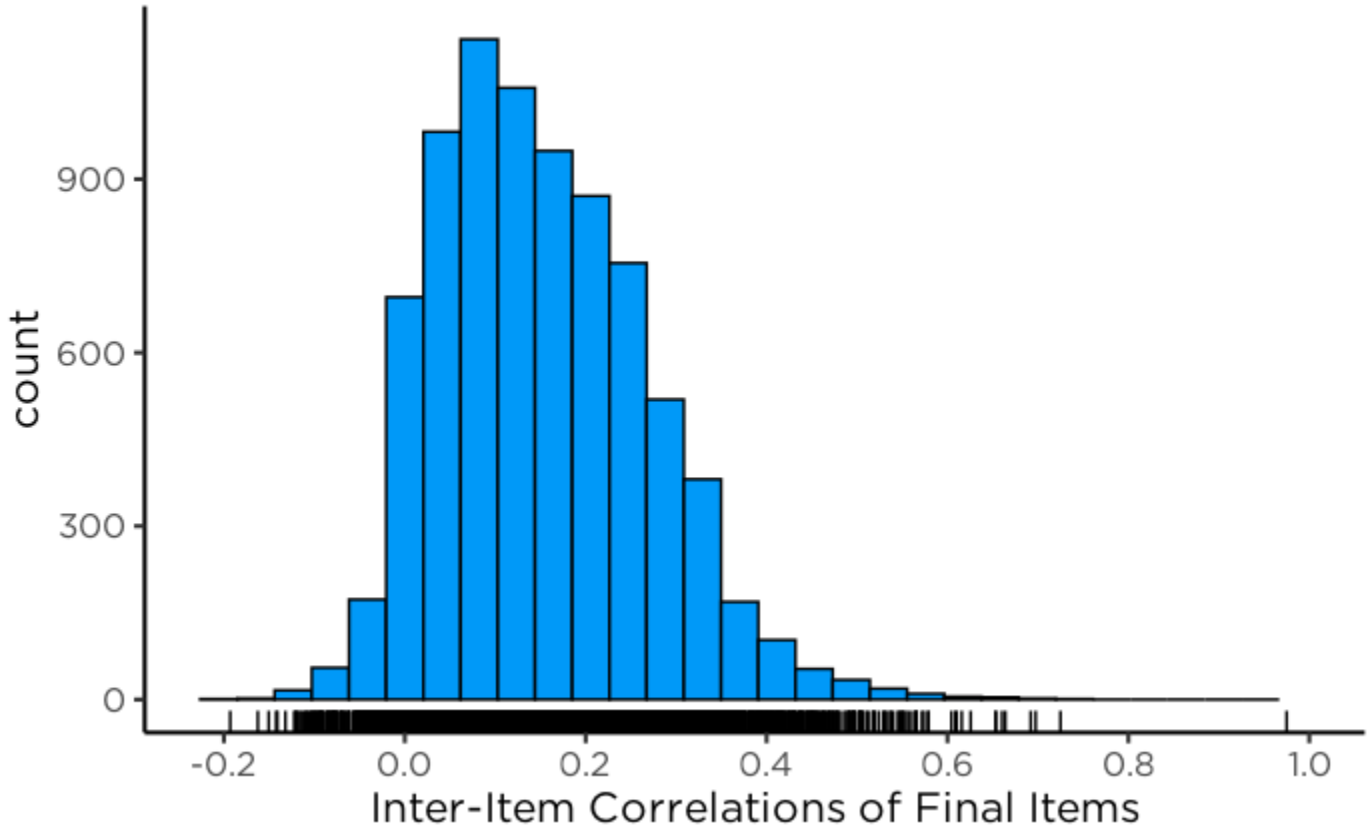
Supplementary Figure S5

Example Questionnaire Item to Assess Absolute Frequency Count of Externalizing Behavior Per Unit Time



Supplementary Figure S6

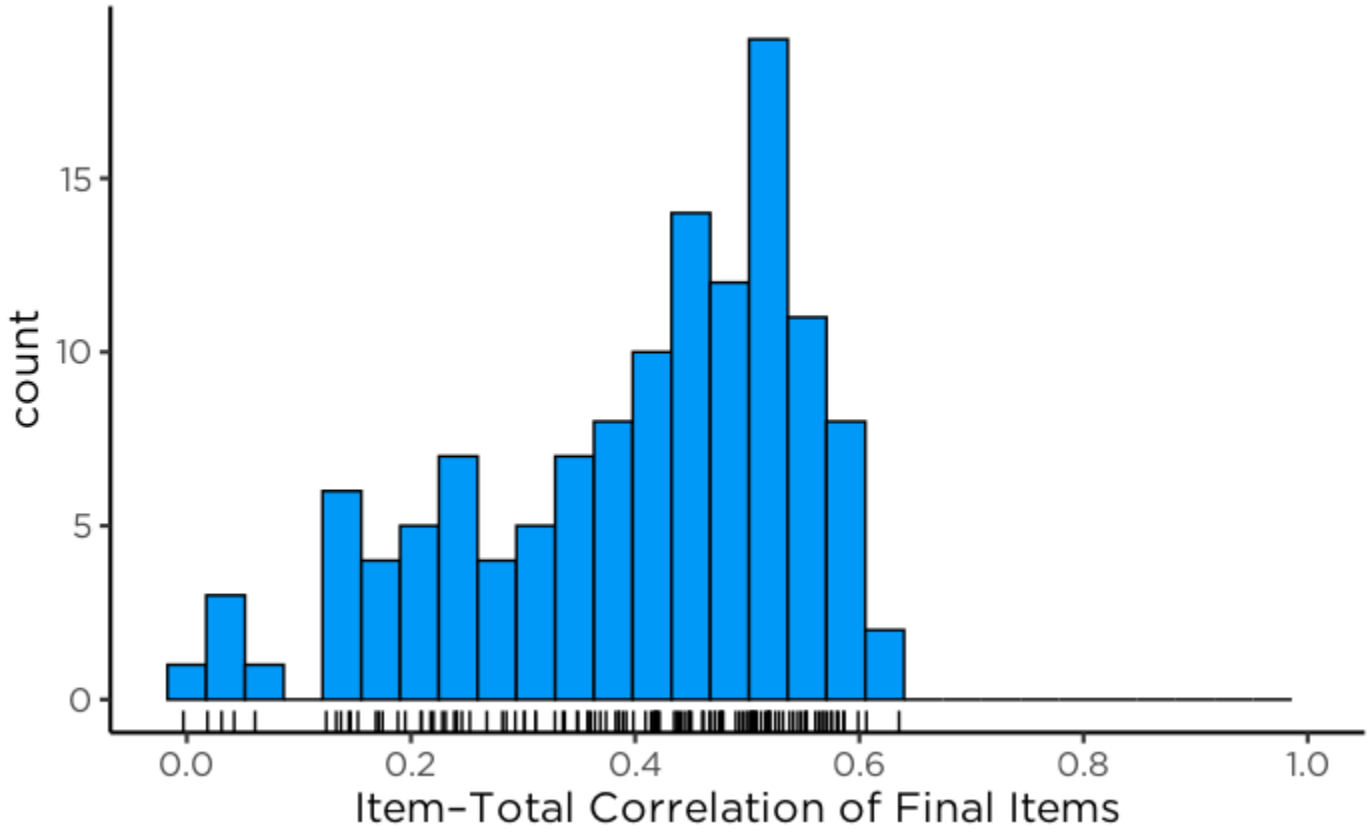
Histogram of Inter-Item Correlations of Final Items



Note. A rug plot is depicted with black lines below the histogram and along the x-axis.

Supplementary Figure S7

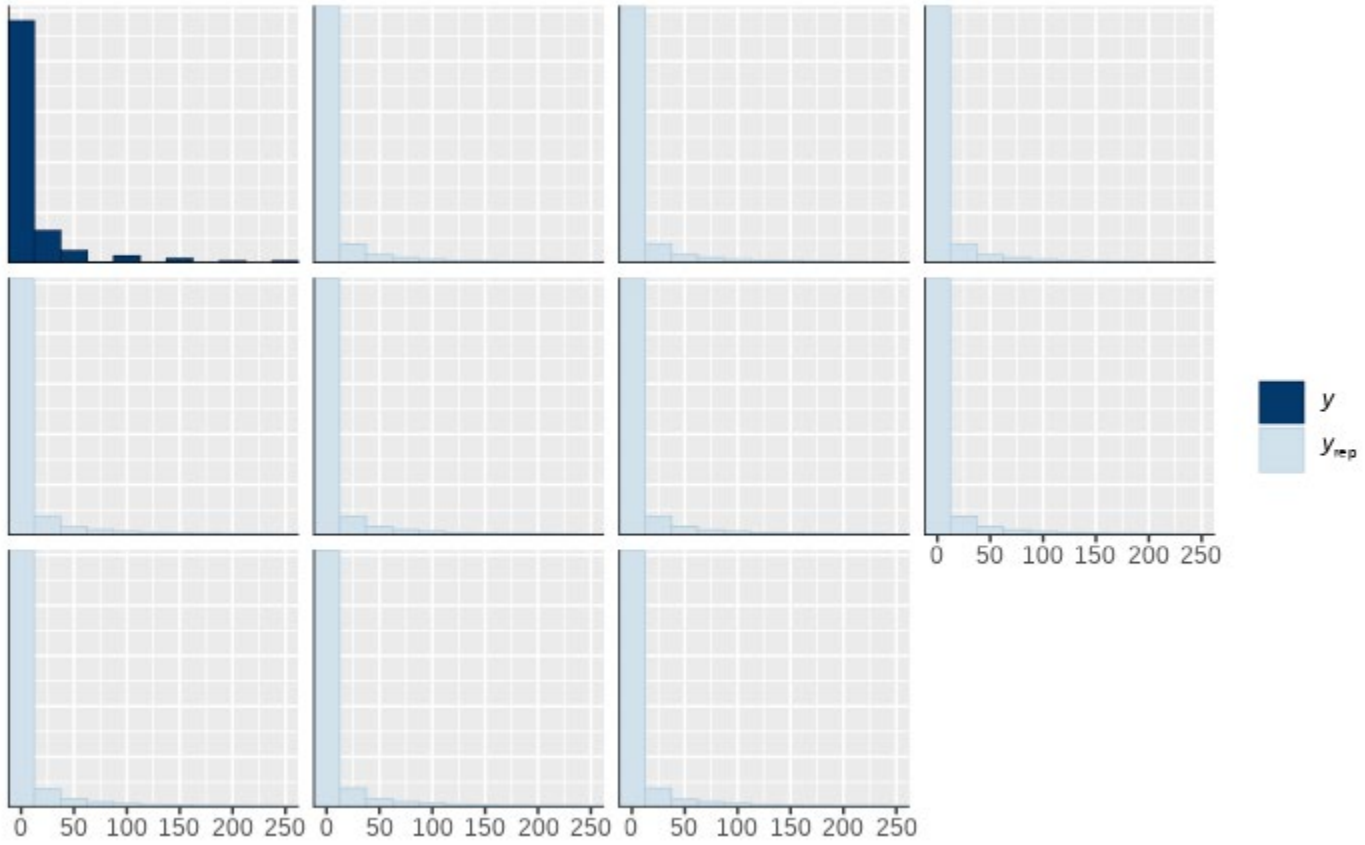
Histogram of Item–Total Correlation of Final Items



Note. A rug plot is depicted with black lines below the histogram and along the x-axis.

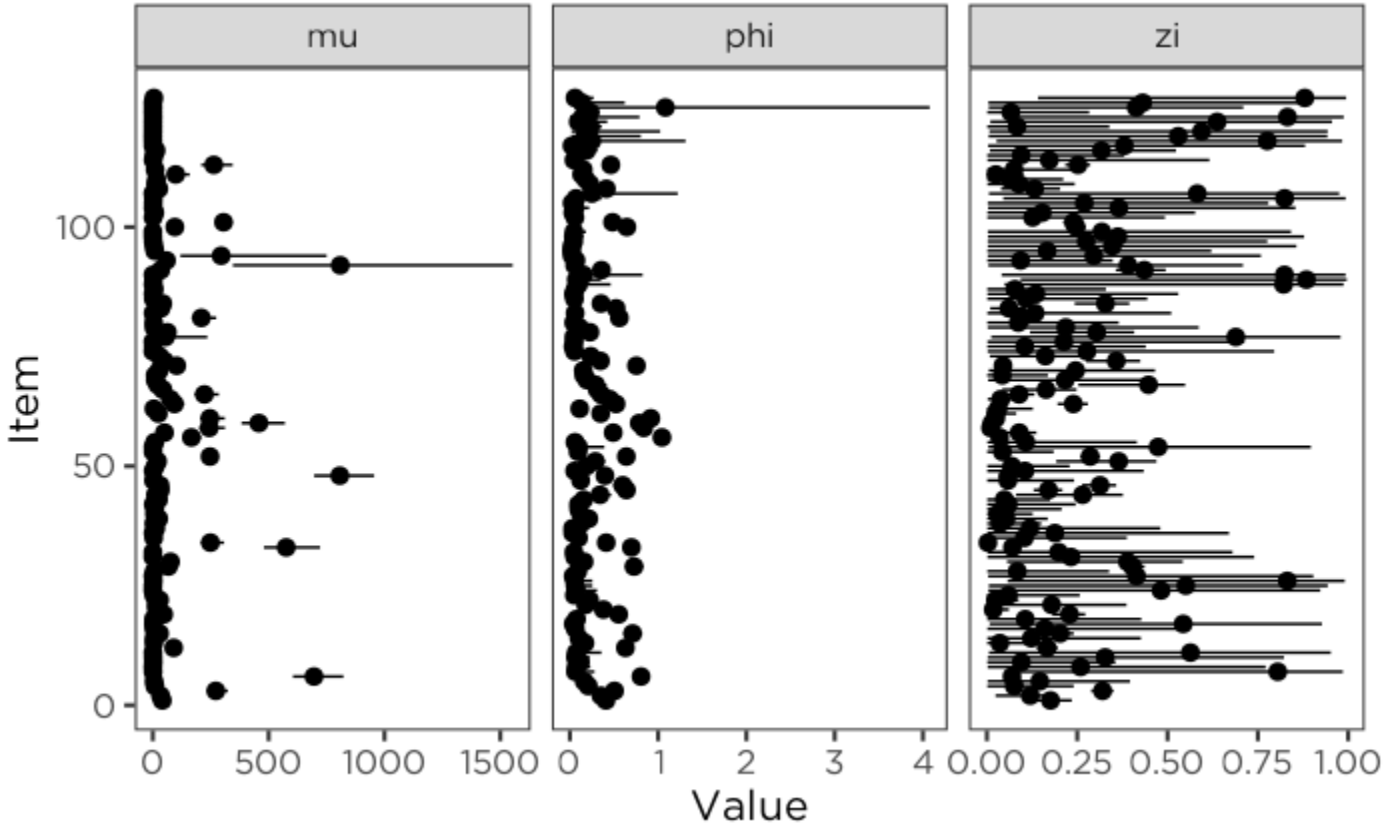
Supplementary Figure S8

Bayesian Item Response Model: Posterior Predictive Check



Supplementary Figure S9

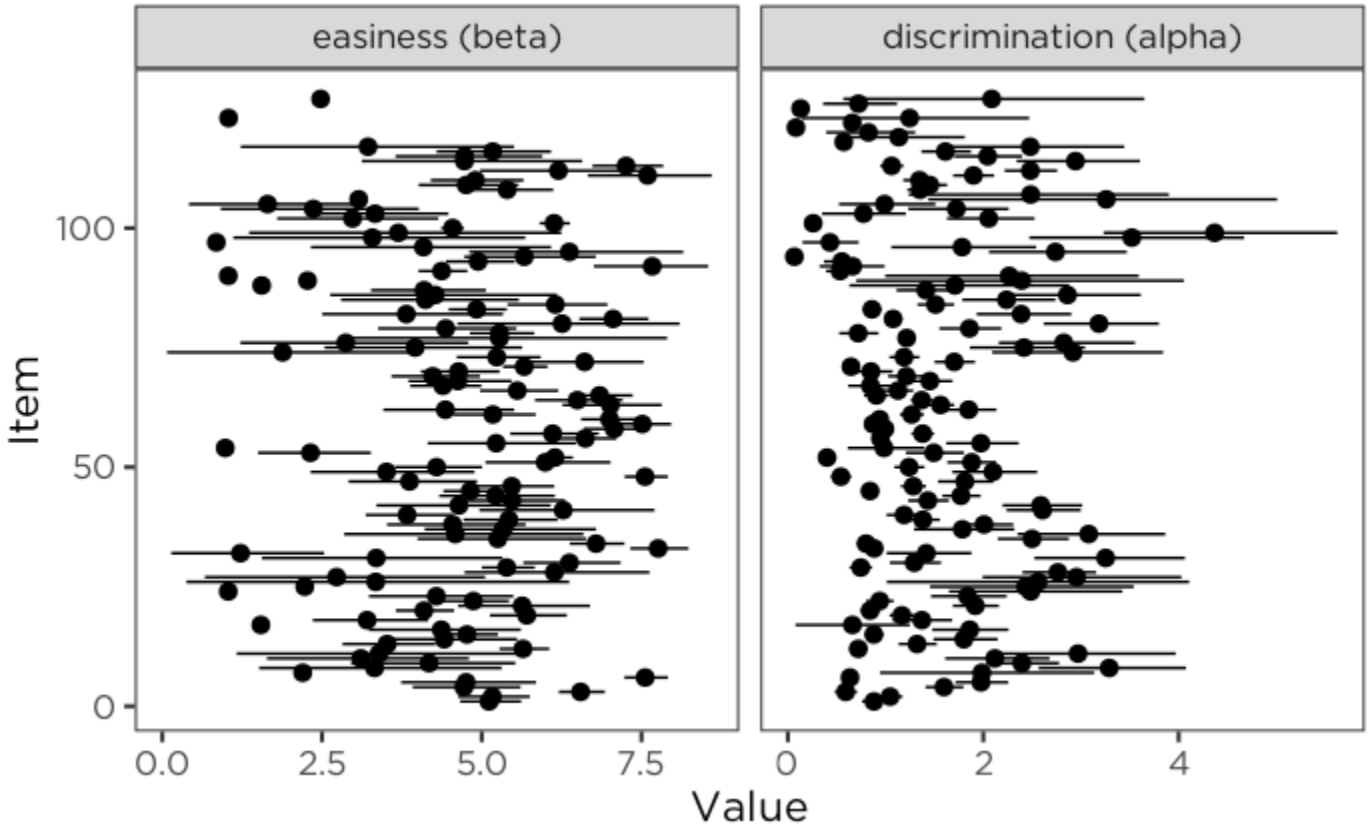
Item Parameters of Bayesian Item Response Model (Zero-Inflated Negative Binomial Model): mu, phi, and ZI



Note. Lines reflect the 95% quantile interval.

Supplementary Figure S10

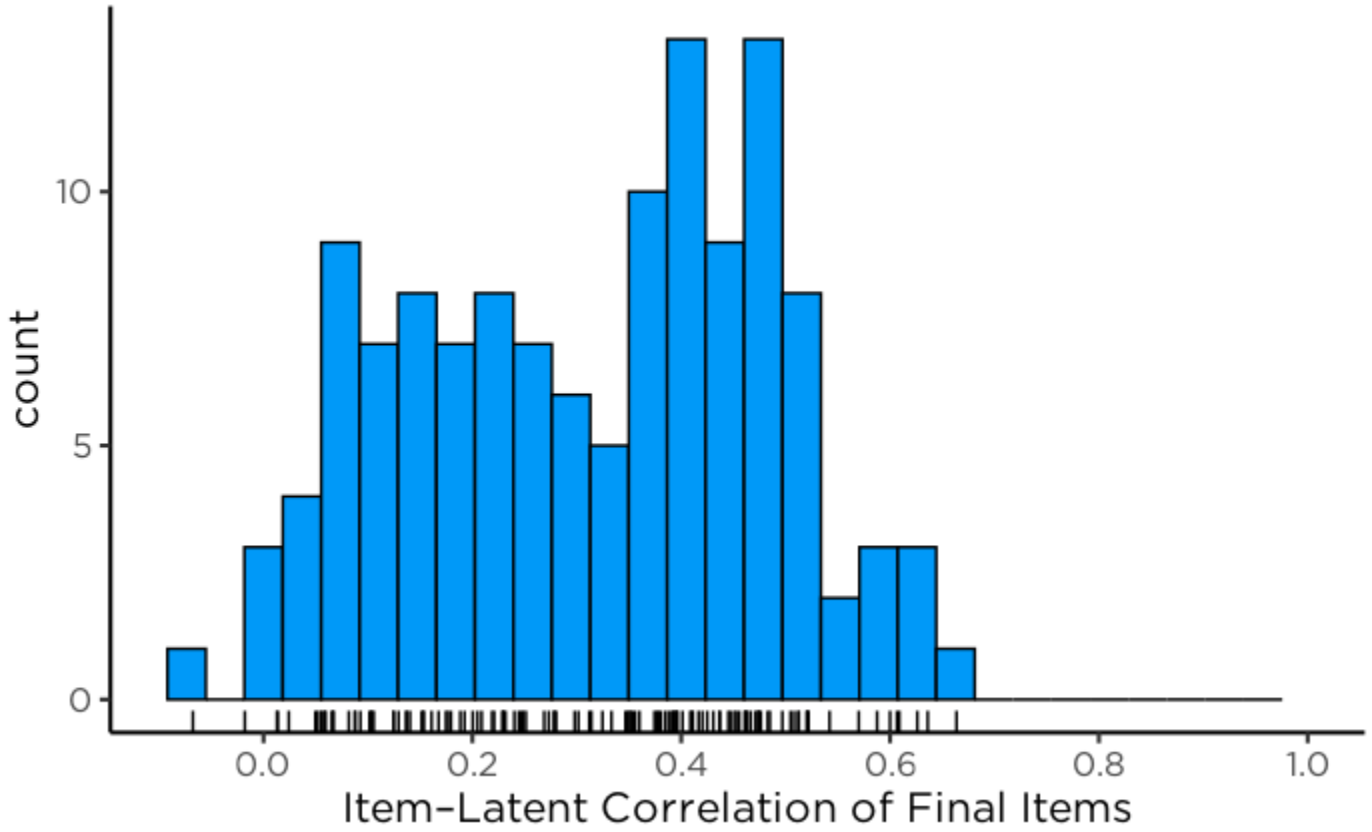
Item Response Theory Parameters of Items From Bayesian Item Response Model: Easiness and Discrimination



Note. Lines reflect the 95% quantile interval.

Supplementary Figure S11

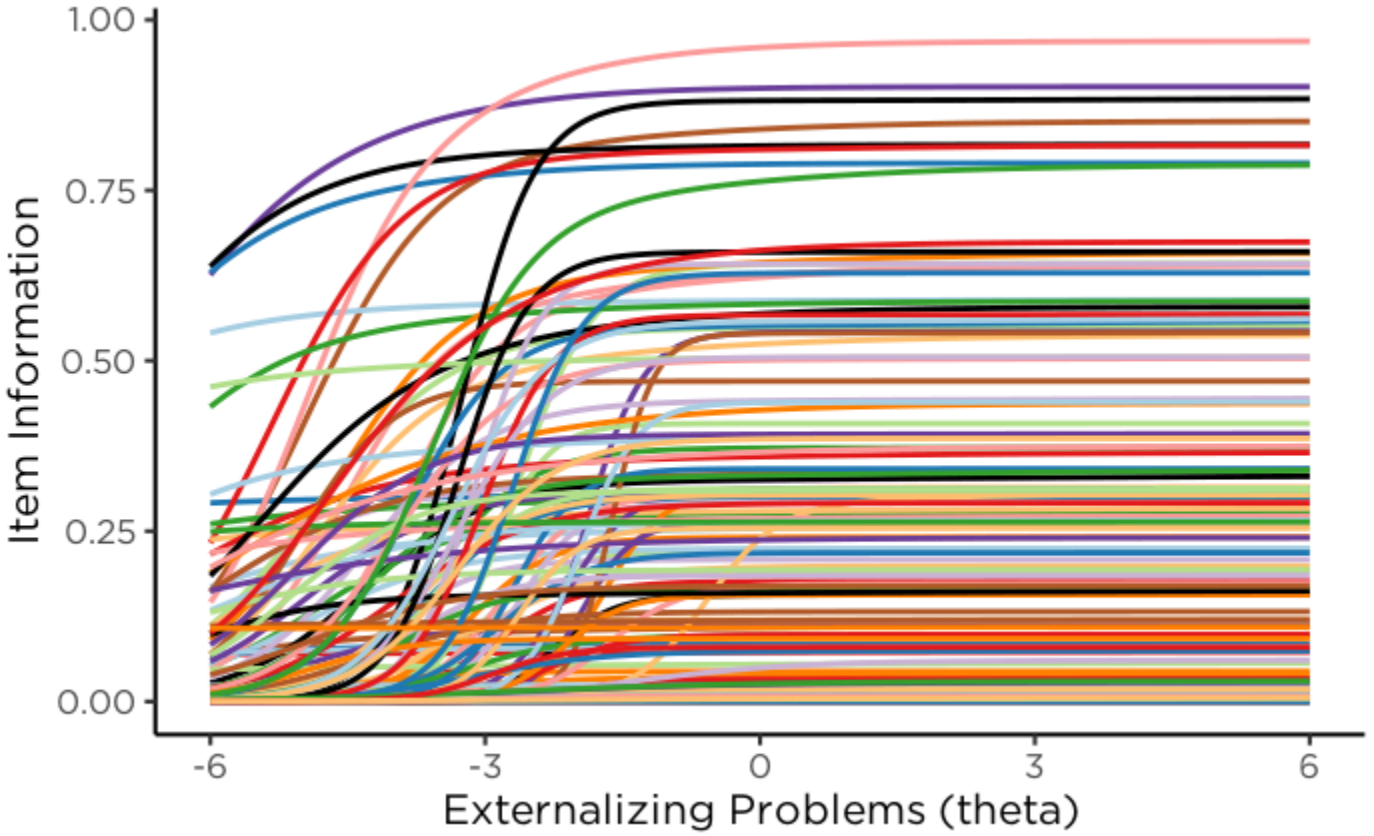
Histogram of Item–Latent Correlation of Final Items



Note. A rug plot is depicted with black lines below the histogram and along the x-axis.

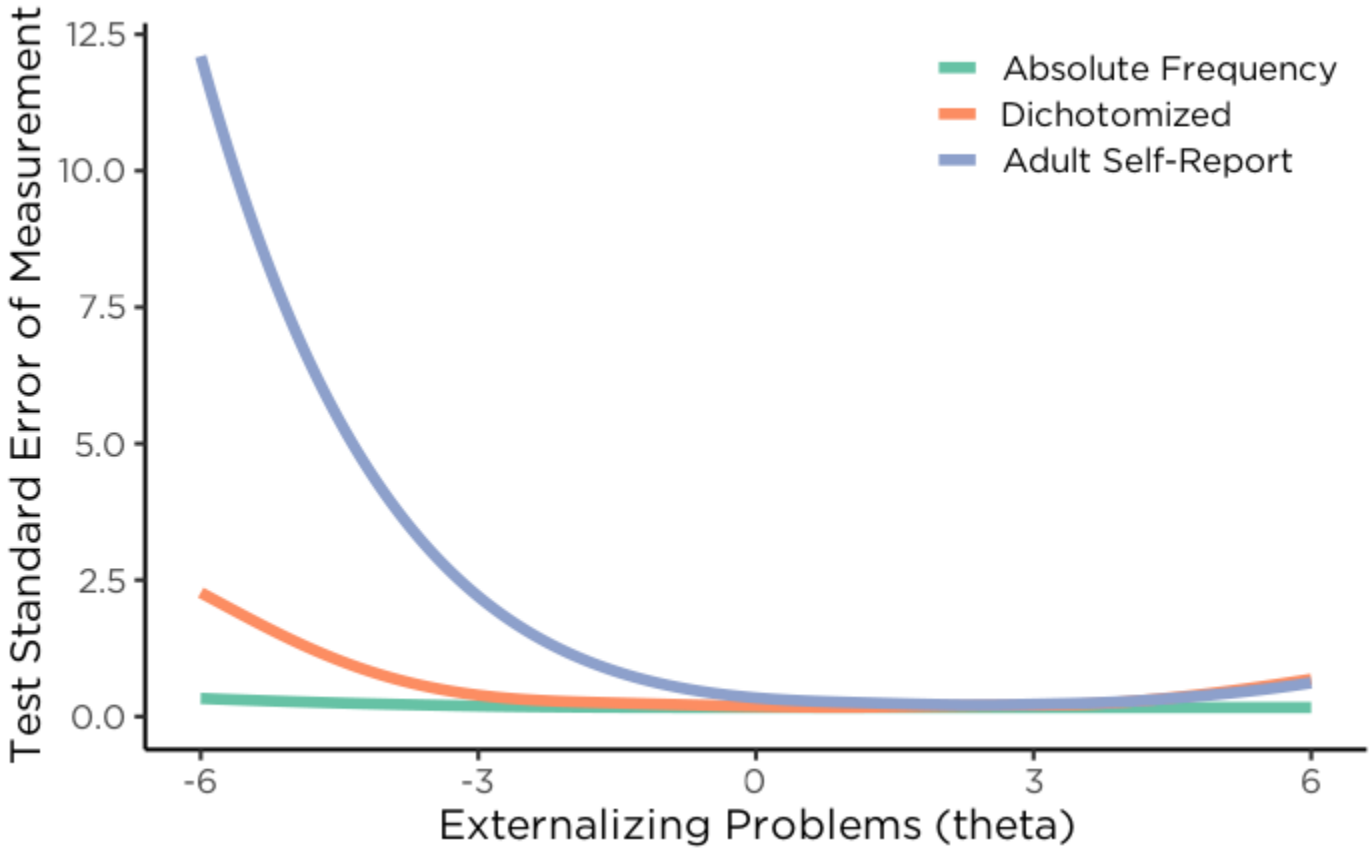
Supplementary Figure S12

Item Information Functions of Items From Bayesian Item Response Model: Item Information as a Function of theta



Supplementary Figure S13

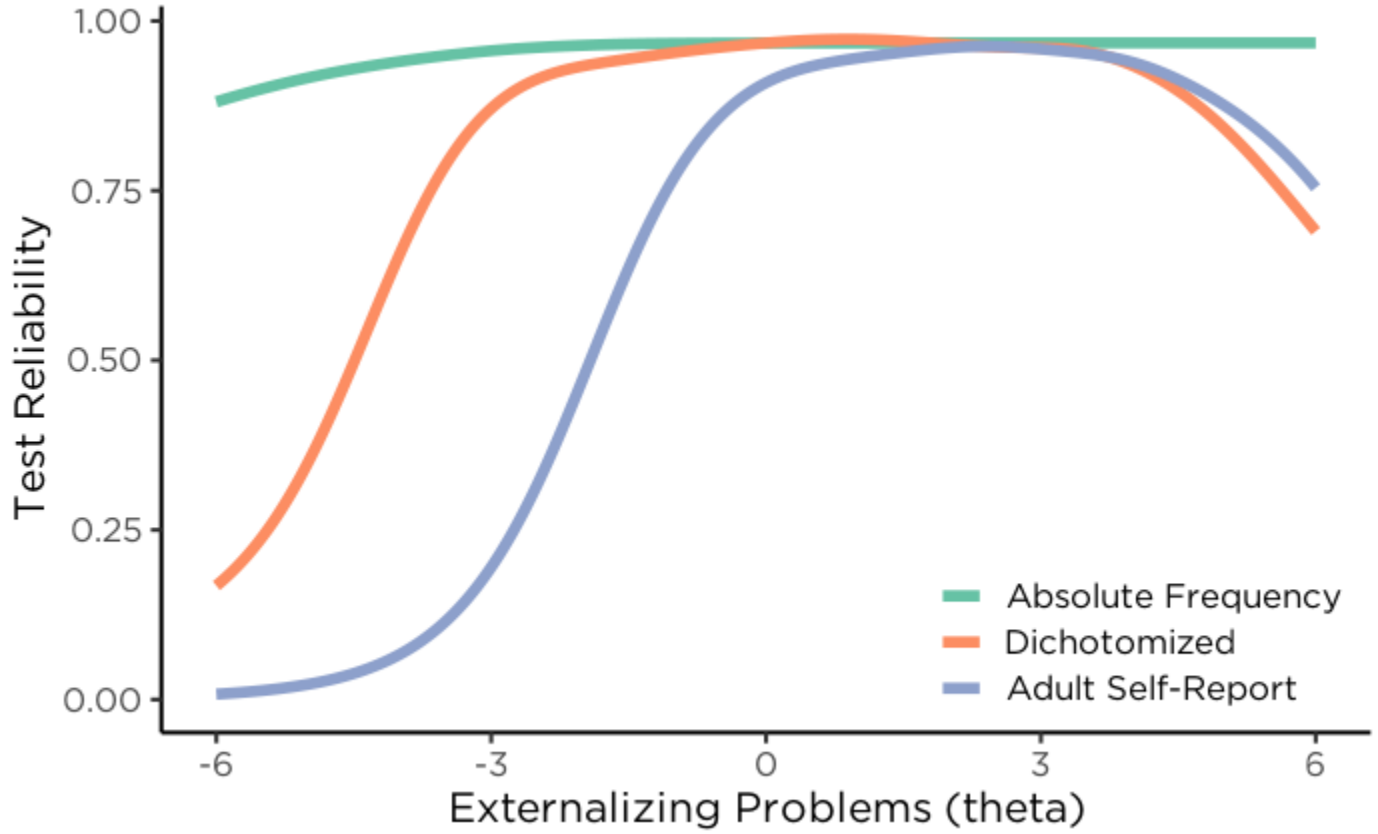
Test Standard Error of Measurement of Absolute Frequency Versus Dichotomized Versus Adult Self-Report Items: Test Standard Error of Measurement as a Function of theta



Note. "Absolute Frequency" refers to count items from the Bayesian item response model; "Dichotomized" refers to dichotomized versions of the count items that were fit to an item response theory model; "Adult Self-Report" refers to items from the Adult Self-report that were fit to an item response theory model. "theta" represents the person's level on the latent externalizing problems factor.

Supplementary Figure S14

Test Reliability of Absolute Frequency Versus Dichotomized Versus Adult Self-Report Items: Test Reliability as a Function of theta



Note. "Absolute Frequency" refers to count items from the Bayesian item response model; "Dichotomized" refers to dichotomized versions of the count items that were fit to an item response theory model; "Adult Self-Report" refers to items from the Adult Self-report that were fit to an item response theory model. "theta" represents the person's level on the latent externalizing problems factor.