Advancing the Measurement of Adolescent Antisocial Behaviour: A Systematic Review and Psychometric Validation of the Antisocial Behaviours Questionnaire (ABQ)

Punita Panchal

Candidate Registration Number: 100458861

Primary Supervisor: Dr Kenny Chiu

Secondary Supervisor: Dr Sharon A S Neufeld

A thesis submitted in partial fulfilment of the degree of

Doctorate in Clinical Psychology

University of East Anglia

Faculty of Medicine and Health Sciences

Norwich Medical School

Submission: 3rd March 2026

Thesis portfolio word count: 26,670 (excluding title page, copyright statement, access condition & agreement, table of contents, list of tables, list of figures, acknowledgements, supplementary material and appendices)

This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with the author and that use of any information derived therefrom must be in accordance with current UK Copyright Law. In addition, any quotation or extract must include full attribution.

Thesis Portfolio Abstract

Background: Precise measurement of adolescent antisocial behaviour (AB) is essential for effective assessment and intervention. However, instruments vary in taxonomic scope and psychometric robustness. **Aims:** This thesis aimed to (1) systematically evaluate adolescent AB instruments, and (2) assess the psychometric properties of the Antisocial Behaviours Questionnaire (ABQ), a brief measure. **Methods:** A systematic review was conducted in accordance with PRISMA and COSMIN guidelines. A 12-tier taxonomic audit mapped item content to the four DSM-5 Conduct Disorder (CD) domains. Reliability generalisation meta-analysis was conducted for instrument families represented in ≥ 3 independent studies. The ABQ was evaluated in two UK community cohorts (ROOTS, N = 1,153; NSPN, N = 2,378; ages 14–24). Analyses included exploratory and confirmatory factor analysis using weighted least squares estimation, reliability estimation, measurement invariance testing, and evaluation of construct and predictive validity. **Results:** From 3,353 records, 275 studies representing 125 unique instrument variants were identified. Only 13 variants (10%) demonstrated both moderate-to-high psychometric certainty and full alignment with all four CD domains. Five instrument families met criteria for recommendation based on structural validity and internal consistency. The ABQ demonstrated a unidimensional structure with excellent fit and strong reliability (ω = .93). Measurement invariance was supported across gender and age. Construct validity was demonstrated through convergence with psychopathic traits and prospective prediction of substance misuse. A threshold of ≥ 2 behaviours optimised diagnostic utility for Conduct Disorder. **Conclusions:** DSM-aligned and psychometrically supported measures remain limited. This portfolio provides empirical support for the ABQ as a concise, DSM-aligned tool for high-precision

screening of antisocial behaviour. **Keywords:** antisocial behaviour; adolescents; psychometrics; systematic review; confirmatory factor analysis; DSM-5; measurement invariance.

Access Condition and Agreement

Each deposit in UEA Digital Repository is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the Data Collections is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form. You must obtain permission from the copyright holder, usually the author, for any other use. Exceptions only apply where a deposit may be explicitly provided under a stated licence, such as a Creative Commons licence or Open Government licence.

Electronic or print copies may not be offered, whether for sale or otherwise to anyone, unless explicitly stated under a Creative Commons or Open Government license. Unauthorised reproduction, editing or reformatting for resale purposes is explicitly prohibited (except where approved by the copyright holder themselves) and UEA reserves the right to take immediate 'take down' action on behalf of the copyright and/or rights holder if this Access condition of the UEA Digital Repository is breached. Any material in this database has been supplied on the understanding that it is copyright material and that no quotation from the material may be published without proper acknowledgement.

Table of Contents

Thesis Portfolio Abstract **2**

Table of Contents **5**

List of Tables **9**

List of Figures **12**

Acknowledgements **13**

Chapter 1. Introduction to the Thesis Portfolio **23**

1.1 Antisocial Behaviour Across Development 23

1.2 Antisocial Behaviour Within the Externalising Spectrum 24

1.3 Challenges in Measuring Antisocial Behaviour 26

1.3 The Need for a Systematic Evaluation of Adolescent Antisocial Behaviour Measures 28

1.4 Evaluating Concise Measures Within Contemporary Psychometric Standards 30

1.5 Empirical Evaluation of the Antisocial Behaviour Questionnaire 31

1.6 Aims of the Thesis Portfolio 33

Chapter 2. A Systematic Review and Meta-Analysis of Adolescent Antisocial Behaviour Scales **35**

Abstract *36*

Introduction *37*

The Present Review: Aim and Contribution 40

Methods *41*

Systematic Review Protocol and Registration 41

Information Sources and Search Strategy 41

Eligibility (Inclusion and Exclusion) Criteria 42

Quality and Psychometric Assessment 43
Statistical Synthesis and Reliability Generalisation 44
Results *45*
Study Selection 45
Description of Studies and Instruments 45
Methodological Quality of the Evidence Base 46
Synthesis of Results: Quality of the Instruments 48
Final Recommendations and Categorization 51
Discussion *52*
Principal Findings 52
Quality of the Measurement Instruments 53
Methodological Quality of the Evidence Base 54
Strengths and Limitations 56
Conclusions 57
References *59*
Supplementary Materials *75*

Chapter 3. Bridging Chapter: From Systematic Appraisal to Empirical Evaluation 103

Chapter 4. Psychometric Evaluation, Diagnostic Utility, and Measurement Invariance of the Antisocial Behaviours Questionnaire (ABQ) in two UK Adolescent Cohorts 106

Abstract *107*
Introduction *108*
Method *112*
Participants and Data Sources 112
Measures 113

Data Preparation and Missingness 114
Statistical Analyses 116
Factor Structure (EFA/CFA) 116
Measurement Invariance 117
Reliability 119
Construct Validity 119
Diagnostic Utility (ROC) 120
Ethical Considerations 121
Results *122*
Descriptive and Preparation Checks 122
Factor Structure 122
Measurement invariance 124
Reliability 126
Diagnostic Utility (ROC) 126
Construct Validity 127
Discussion *129*
The ABQ: Psychometric Foundation and Developmental Stability. 129
Strengths and Limitations 131
Clinical Implementation 132
References *134*
Supplementary Materials *149*
Chapter 5. Discussion and Critical Analysis 169
Summary of Principal Findings 169
Integrated Contribution of the Portfolio 171
Defining and Delimiting Antisocial Behaviour for Measurement 172
Psychometric Standards and Developmental Comparability 174

Feasibility, Informant Perspectives, and Applied Use 178

Critical Evaluation of the Portfolio 179

Implications and Future Directions 181

Conclusion 182

References 184

Appendices 189

Appendix A: Confirmation of Full Ethical Approval from FMH REC 189

List of Tables

Systematic Review

Table 1. Landscape of Assessment Instruments and Sample Characteristics Grouped by Taxonomic Category 60

Table 2. Directory and Structural Characteristics of Included Measurement Instruments 61

Table 3. COSMIN Synthesis: Methodological Quality and Evidence Certainty by Instrument Family 62

Table 4. Operational Directory of High-Alignment (Five Instrument Families) Diagnostic Instruments 63

Table 5. Reliability Generalisation Meta-Analysis Results Categorised by Taxonomic Alignment 64

Systematic Review Supplementary Materials

Table S1. Electronic Database Search Strategy: Concept Blocks, Controlled Vocabulary, and Free-Text Terms 67

Table S2. Eligibility Criteria for Study Selection and Operational Decision Codes 68

Table S3. Taxonomy of Antisocial Construct Categories and Inclusion Rationale 69

Table S4. List of Studies Excluded at the Full-Text Assessment Stage with Rationales 71

Table S5. Studies Excluded During the Consolidation Phase: Population Age Refinement (n=69) 79

Table S6. Data Extraction Framework: Parameter Definitions and Psychometric Variables 82

Table S7. Evaluative Criteria for Rating the Quality of Measurement Property Outcomes 84

Table S8. Full Psychometric Extraction and Methodological Quality for 125 Included Instrument Variants 85

Empirical Project

Table 1. Antisocial Behaviours Questionnaire (ABQ): Item Stems and Content Summaries 135
Table 2. Confirmatory Factor Analysis 136
Table 3. Measurement Invariance Across Gender 137
Table 4. ROC operating characteristics at selected thresholds (ROOTS Baseline) 138
Table 5. Key Convergent, Divergent, and Predictive Validity of the ABQ Across Cohorts and Waves 139

Empirical Project Supplementary Materials

Table S1. Sample Characteristics and Retention by Cohort and Wave 141
Table S2. Overview of Samples and Analyses 142
Table S3. Comparator measures used in validity analyses: subscales, response formats, waves, and role 143
Table S4. ABQ Item Content and Baseline Endorsement (%) in ROOTS and NSPN Cohorts 144
Table S5. Tetrachoric correlations among ABQ items (ROOTS df1, n = 577; final 8-item set) 145
Table S6. Tetrachoric correlations among ABQ items (ROOTS df2, n = 576; final 8-item set) 146
Table S7. Tetrachoric correlations among ABQ items (NSPN 14–15 years, n ≈ 487; final 8-item set) 147
Table S8. Exploratory Factor Analysis of the ABQ (ROOTS df1, n = 577; final 8-item set) 150
Table S9. Bootstrapped 95% Confidence Intervals for Standardised Residual Variances for the One-Factor ABQ Model 151
Table S10. Gender Measurement Invariance of the ABQ Across ROOTS and NSPN 152
Table S11. Measurement Invariance Across Age bands 153
Table S12. Reliability and Factor Adequacy of the 8-item ABQ (N = 1,153) 154
Table S13. Item-Level Reliability Diagnostics for the ABQ (ROOTS N = 1,153) 155
Table S14. Full ROC Threshold Grid for ABQ Predicting K-SADS Conduct Disorder (Baseline ROOTS) 156

Table S15. Full ABQ Convergent, Divergent, and Predictive Validity Across Cohorts and Waves 157

List of Figures

Systematic Review

Figure 1. PRISMA 2020 Flow Diagram of the Systematic Search and Selection Process for Adolescent Antisocial Behaviour Instruments 65

Figure 2. Distribution of Methodological Quality (Risk of Bias) Ratings for Measurement Properties across 125 Instrument Variants 66

Systematic Review Supplementary Materials

Figure S1. Forest Plot of Pooled Internal Consistency Coefficients for Key Adolescent Antisocial Behaviour Instrument Families..... 93

Empirical Project Supplementary Materials

Figure S1. Parallel Analysis Plot for the ABQ (ROOTS splithalf1) 148

Figure S2. Scree Plot of Eigenvalues for ABQ (ROOTS splithalf1) 149

Figure S3. ABQ-Comparator Correlations Across Cohorts and Waves..... 161

Acknowledgements

I would like to express my sincere gratitude to my primary supervisor, Dr Kenny Chiu, for his leadership and guidance throughout this project. His critical insight, strategic perspective, and steady encouragement provided an essential framework within which this work developed. I am particularly grateful for his support at key stages of the research process and for the confidence he placed in me as the project evolved.

I am especially grateful to my secondary supervisor, Dr Sharon A. S. Neufeld, whose sustained involvement and methodological expertise were instrumental to this work. Her consistent weekly supervision, detailed feedback, and thoughtful troubleshooting at each stage of the project provided both intellectual clarity and practical direction. Her generosity with her time, depth of statistical knowledge, and commitment to refining the work have had a lasting impact on both this portfolio and my development as an independent researcher.

I am grateful to the research teams responsible for the ROOTS and NSPN studies for granting access to these longitudinal datasets, and to the young people and families whose participation continues to inform research.

I would also like to thank the Clinical Psychology programme team at the University of East Anglia for fostering intellectual curiosity and methodological rigour. The research training provided across the programme has been foundational to this work.

To my cohort, thank you for the solidarity, encouragement, and shared perseverance that characterised our training. The sense of mutual support throughout this journey has been both grounding and sustaining.

To my friends, thank you for the encouragement, perspective, and steadiness you have offered throughout this journey. Your willingness to listen, to celebrate milestones, and to provide balance beyond academia has been deeply appreciated.

Finally, I am deeply grateful to my family for their unwavering belief in me, even at times when my own confidence faltered. Their patience, understanding, and steadfast encouragement have carried me through the challenges and uncertainties of this training. They have been a constant source of strength, perspective, and reassurance, and this achievement is as much theirs as it is mine.

Glossary of Terms and Abbreviations

The following glossary defines the key terms and lists the abbreviations and statistical symbols used throughout this thesis portfolio. Terms and abbreviations are listed together in a single alphabetical sequence; statistical notation follows.

Terms and abbreviations

Term / Abbreviation	Definition
AB	Antisocial behaviour — actions that violate social norms, legal rules, or the rights of others (rule-breaking, deceit, aggression, property violation), spanning a continuum from minor transgressions to serious harm.
ABQ	Antisocial Behaviours Questionnaire — a concise (8-item) DSM-aligned self-report measure of antisocial acts.
ACEs	Adverse Childhood Experiences — early adversities treated as a distal correlate of antisocial behaviour and screened out by the taxonomic audit.
ADHD	Attention-Deficit/Hyperactivity Disorder — a regulatory/attentional disorder excluded from the antisocial construct by the taxonomic audit.
Age–crime curve	The robust developmental pattern whereby antisocial/criminal behaviour rises in adolescence, peaks in the late teens, and declines into adulthood.
AL	Adolescence-limited — antisocial behaviour confined to adolescence (Moffitt's developmental taxonomy; contrast LCP).
APA	American Psychological Association.
APSD	Antisocial Process Screening Device — a youth measure of psychopathic traits.
ASB	Antisocial behaviour — see AB, the abbreviation used throughout this portfolio.
ASD	Autism Spectrum Disorder — a neurodevelopmental condition excluded from the antisocial construct by the taxonomic audit.
ASEBA	Achenbach System of Empirically Based Assessment — a multi-informant family of measures (e.g., CBCL, YSR, TRF).
ASPD	Antisocial personality disorder — the adult diagnostic outcome to which persistent childhood/adolescent antisocial symptoms may contribute.
AUC	Area under the (ROC) curve — a summary of overall discrimination (.50 = chance; higher is better).
Bartlett's test of sphericity	A test of whether the correlation matrix is factorable (i.e., not an identity matrix).
BASC	Behavior Assessment System for Children (third edition, BASC-3) — a broadband clinical assessment.
Base rate / prevalence	The proportion of true cases in the population, which influences predictive values.
BCS-A	Bullying and Cyberbullying Scale for Adolescents.

Behavioural phenotype	The observable behavioural expression of an underlying trait, disposition, or liability (here, the antisocial phenotype).
Behavioural topography	Item-level mapping of an instrument's content onto predefined antisocial behavioural domains, looking past subscale labels.
Bifactor model	A model specifying a general factor alongside specific factors; known often to yield favourable fit indices.
Binarised / binary indicator	An item recoded into two categories (e.g., any endorsement vs none).
BIS-11	Barratt Impulsiveness Scale-11 — a self-report measure of impulsivity.
Bootstrap confidence interval	A confidence interval derived by repeatedly resampling the data to assess parameter stability.
BPAQ-SF	Buss-Perry Aggression Questionnaire – Short Form (an aggression measure noted in the review).
Broadband externalising measure	An instrument indexing the wide externalising spectrum (multiple co-occurring problem behaviours) rather than a single, specific behaviour.
CADS	Child and Adolescent Dispositions Scale.
CAMHS	Child and Adolescent Mental Health Services.
CBCL	Child Behavior Checklist — caregiver-report measure within the ASEBA family.
CD	Conduct disorder — a DSM-5 diagnosis for minors defined by a persistent pattern of behaviour violating others' rights or major age-appropriate norms (≥ 3 of 15 symptoms with functional impairment), across four domains.
CFA	Confirmatory factor analysis — a hypothesis-driven method that tests a pre-specified factor structure.
CFI	Comparative Fit Index — an incremental model-fit index (≥ .95 indicates good fit).
CI	Confidence interval — a range expressing the precision/uncertainty of an estimate.
Classical test theory	The traditional measurement framework treating an observed score as a true score plus measurement error.
Communality	The proportion of an item's variance explained by the retained factors.
Concurrent validity	Criterion validity assessed when the measure and the criterion are collected at the same time.
Conduct pathology	Clinically significant, persistent antisocial behaviour, as distinct from normative transgression.
Configural invariance	Equivalence of the overall factor structure (the same pattern of loadings) across groups.
Congeneric model	A measurement model allowing items to have unequal factor loadings (contrast with a tau-equivalent model).
Construct alignment	The degree to which instrument items map onto theoretically defined behavioural domains; a pragmatic proxy for validity where no gold standard exists.
Construct / content coverage	The extent to which a measure's items represent the full range of the target behavioural domain.

Construct drift	The gradual broadening or blurring of a construct so that a measure captures content beyond its intended scope.
Construct validity	Cumulative evidence, from multiple sources, that a measure reflects the intended theoretical construct.
Convergent validity	Expected positive associations between a measure and indicators of related constructs.
Correlated-factor model	A model comprising multiple correlated first-order factors.
Correlated residuals	Modelled covariance between item error terms, often introduced via post-hoc modification.
COSMIN	COnsensus-based Standards for the selection of health Measurement INstruments — a standardised methodology for evaluating instruments' measurement properties.
CPAQ-R	Child-to-Parent Aggression Questionnaire (Revised).
Criterion validity	The correspondence of a measure with an external criterion (e.g., a diagnostic benchmark).
Cross-loading	An item that loads substantially on more than one factor.
CTV	Child-to-Parent/Caregiver Violence.
Cut-off / threshold	The score at or above which an individual is classified as a likely case.
CU traits	Callous-unemotional traits — lack of remorse/guilt, shallow affect, callous lack of empathy, and unconcern about performance (basis of the DSM-5 “limited prosocial emotions” specifier).
DASI	Drug, Alcohol and Self-Injury Scale.
DClinPsy	Doctorate in Clinical Psychology.
Developmental homogeneity	Restriction of a sample to a comparable developmental range so that findings are not confounded by age.
Developmental psychopathology	The study of the origins, course, and variation of maladaptation and disorder across development.
Developmental trajectory	The course of antisocial behaviour over time, which may follow a transient or a persistent pathway.
Diagnostic utility	A measure's ability to discriminate between individuals with and without a given diagnosis.
DIF	Differential item functioning — when respondents at the same trait level but in different groups respond differently to an item.
Dimensionality	The number of latent factors required to account for item responses (e.g., single- vs multi-factor).
Dimensional vs categorical models	Representations of psychopathology as continuous dimensions versus discrete diagnostic categories; AB is dimensional, whereas CD is categorical.
Disinhibition	A general propensity toward poorly controlled, impulsive, norm-violating behaviour thought to underlie the externalising spectrum.
Distal correlate	A factor associated with antisocial behaviour but not itself antisocial behaviour (e.g., trauma, adverse childhood experiences); excluded from the construct definition.
Divergent (discriminant) validity	Expected weak or negative associations between a measure and indicators of conceptually distinct constructs.

DSM-5	Diagnostic and Statistical Manual of Mental Disorders (Fifth Edition).
DSM-IV	Diagnostic and Statistical Manual of Mental Disorders (Fourth Edition).
DWLS	Diagonally Weighted Least Squares — an estimator for categorical/ordinal indicators.
ECV	Explained Common Variance — the proportion of common variance attributable to a general factor (ECV = 1.00 indicates unidimensionality).
EFA	Exploratory factor analysis — a data-driven method to identify the number and nature of latent factors.
Effect size	A standardised index of the magnitude of an association or difference.
Eigenvalue	The amount of variance accounted for by a factor.
EJPA	European Journal of Psychological Assessment.
Emerging adulthood	A developmental period (approximately 18–25 years) recognised as a prolonged transition to adult psychosocial functioning.
EP	Empirical Paper (empirical project).
ESEM	Exploratory structural equation modelling — a hybrid allowing cross-loadings within a confirmatory framework.
Essential unidimensionality	The condition in which a dominant general factor justifies use of a single total score despite minor multidimensionality.
Externalising spectrum	A broad higher-order dimension of outwardly directed difficulties (aggression, impulsivity, rule-breaking, disinhibition), within which antisocial behaviour is situated.
Face validity	The superficial appearance that a measure assesses what it claims; insufficient on its own to establish coverage.
Factor analysis	A family of statistical methods for identifying or testing the latent structure underlying item responses.
Factor loading	The strength of association between an item and its latent factor.
Factor score	A score estimated on the latent factor, as distinct from a simple summed score.
Four DSM-5 CD domains	The four behavioural categories defining conduct disorder: aggression to people and animals; destruction of property; deceitfulness or theft; and serious violations of rules.
Functional impairment	Clinically significant disruption to everyday functioning, required alongside symptom count for a CD diagnosis.
Gold standard	A definitive criterion (e.g., a structured diagnostic interview) against which a measure is validated; often absent for antisocial behaviour.
GRADE	Grading of Recommendations, Assessment, Development and Evaluations — an approach for rating the overall certainty of a body of evidence.
Grey literature	Non-peer-reviewed sources (e.g., theses, conference abstracts) excluded here to maintain an evidence-level threshold.
HCPC	Health and Care Professions Council.
Heterogeneity	Between-study variability in results beyond that expected by chance (indexed by Cochran's Q and I^2).
Higher-order model	A model in which a superordinate factor accounts for the correlations among lower-order factors.

HiTOP	Hierarchical Taxonomy of Psychopathology — a dimensional model organising psychopathology into empirically derived spectra rather than categorical diagnoses.
ICC	Intraclass correlation coefficient — an index of inter-rater or test–retest reliability.
ICU	Inventory of Callous-Unemotional Traits (ICU-11 denotes the 11-item version).
Incremental fit indices	Indices (e.g., ΔCFI) used to compare nested models against one another.
Internal consistency	The degree to which items intended to measure the same construct correlate with one another (e.g., α, ω).
Internalising	Inwardly directed psychological difficulties such as anxiety and depression; the counterpart to externalising problems.
Inter-rater reliability	Agreement between different raters/informants assessing the same individual (e.g., indexed by the intraclass correlation, ICC).
IRT	Item Response Theory — a modelling framework relating a latent trait to the probability of item responses.
KMO	Kaiser-Meyer-Olkin measure of sampling adequacy — indicates whether data are suitable for factor analysis.
K-SADS	Kiddie Schedule for Affective Disorders and Schizophrenia (Present and Lifetime; K-SADS-PL) — a structured diagnostic interview used here as the gold-standard criterion.
Latent liability / latent construct	An unobserved underlying propensity (e.g., toward antisocial behaviour) inferred from observed item responses.
Latent mean comparison	Comparison of estimated factor means across groups, valid only when scalar invariance holds.
Latent variable modelling	Statistical modelling of unobserved constructs inferred from observed indicators.
lavaan	Latent variable analysis (R package) used for CFA and invariance testing.
LCP	Life-course-persistent — antisocial behaviour beginning early and persisting across the lifespan (Moffitt's developmental taxonomy; contrast AL).
Limited prosocial emotions	The DSM-5 conduct-disorder specifier capturing clinically significant callous–unemotional traits.
Listwise / pairwise deletion	Excluding cases with any missing data (listwise) or only for the specific analysis affected (pairwise).
MAR	Missing at random — missingness explainable by observed variables, permitting principled handling (e.g., multiple imputation).
MARS	Meta-Analysis Reporting Standards (APA) — reporting guidelines for quantitative synthesis.
MCAR	Missing Completely At Random.
MeSH	Medical Subject Headings — the controlled vocabulary used for database searching.
Meta-analysis	The statistical pooling of quantitative results across independent studies.
MI	Measurement invariance — the property that an instrument measures the same construct equivalently across groups or time (also Multiple Imputation in missing-data contexts).

MIA	Mental Health and Social Inadaptation Assessment.
Minimum residual (minres) extraction	A factor-extraction method that minimises the residual correlations.
ML	Maximum likelihood (estimation) — an estimator that maximises the likelihood of the observed data.
Modification indices	Statistics indicating model changes that would improve fit; used cautiously and for diagnostic purposes.
Multi-group CFA	Confirmatory factor analysis estimated simultaneously across groups to test invariance.
Multi-informant assessment	The use of multiple reporters (self, parent, teacher, clinician) to capture behaviour across contexts.
Multiple imputation	Replacing missing values with several plausible values to reflect uncertainty (a MAR-consistent approach).
Narrow / specialised construct	A single, circumscribed facet or context-specific behaviour targeted by a measure (e.g., cyberbullying, dating violence) rather than the broad externalising domain.
Nested models	Models in which one is a constrained version of another, permitting formal comparison.
NHS	National Health Service (UK).
NICE	National Institute for Health and Care Excellence.
NIHR	National Institute for Health and Care Research.
Nomological network	The web of theoretical relationships a construct is expected to have with other constructs, against which construct validity is judged.
Normative experimentation	Transient, developmentally typical rule-testing in adolescence, to be distinguished from persistent conduct pathology.
NSPN	NeuroScience in Psychiatry Network — a UK adolescent/young-adult cohort study.
Oblimin rotation	An oblique rotation method that allows extracted factors to correlate.
ODD	Oppositional defiant disorder — a pattern of angry/irritable mood and defiance toward authority, distinct from rights-violating antisocial conduct.
PA	Psychological Assessment (journal).
Parallel analysis	A factor-retention method comparing observed eigenvalues with those from random data.
Parsimony	Preference for the simplest model that fits the data adequately.
PCA	Principal component analysis — a data-reduction technique that does not model latent constructs (and so does not establish structural validity).
PCL-YV	Psychopathy Checklist: Youth Version — a clinician-rated measure.
p-factor	General psychopathology factor — a general dimension capturing liability shared across mental disorders.
PICO	Population, Intervention, Comparison, Outcome — a framework for structuring review questions and searches.
Positive / negative predictive value (PPV / NPV)	The probability that a positive/negative screen reflects true case/non-case status; dependent on the base rate.
Predictive validity	The degree to which scores forecast a relevant future outcome.

PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses — the reporting guideline for systematic reviews.
pROC	Receiver operating characteristic curve analysis (R package).
PROSPERO	International Prospective Register of Systematic Reviews.
PSCD	Proposed Specifiers for Conduct Disorder — one of the psychometrically robust, DSM-aligned instrument families identified in the review (distinct from the APSD).
psych	Psychometric analysis (R package) used for EFA and reliability.
Psychometrics	The theory and methods of psychological measurement, concerned with the reliability and validity of scores.
Psychopathic traits	Affective–interpersonal characteristics (e.g., callousness, lack of remorse, narcissism, impulsivity) that frequently co-occur with antisocial behaviour but represent dispositions rather than acts.
R	R statistical computing language.
Random-effects model	A meta-analytic model assuming that true effects vary across studies.
Rasch model	A one-parameter item-response model in which item endorsement depends on the respondent's trait level and the item's difficulty.
Rayyan	Systematic review screening tool.
Reactive vs proactive aggression	Aggression occurring in response to perceived provocation/threat (reactive) versus planned, goal-directed aggression (proactive).
Relational aggression	Harm enacted through manipulation of relationships or social standing rather than physical force.
Reliability	The consistency or reproducibility of measurement.
Reliability generalisation meta-analysis	Meta-analytic pooling of reliability coefficients across samples to evaluate their stability across contexts.
RMSEA	Root Mean Square Error of Approximation — an absolute model-fit index (≤ .06 indicates good fit).
RMSR	Root Mean Square Residual — a residual-based index of model fit reported from exploratory factor analysis (distinct from the standardised SRMR).
ROB	Risk of bias — the extent to which methodological limitations may distort a study's findings.
ROC	Receiver operating characteristic (curve) — a plot of sensitivity against 1 − specificity across thresholds.
ROOTS	UK longitudinal adolescent community cohort study (Goodyer et al., 2010).
RPQ	Reactive–Proactive Aggression Questionnaire.
RQ	Research Question.
SBPS	School Bullying Perpetration Scale.
Scalar invariance	Equality of loadings and item intercepts/thresholds across groups, allowing valid comparison of means.
Screening vs diagnostic instrument	Brief tools estimating general liability for triage/monitoring, versus instruments assessing symptoms and impairment for formal diagnosis.

Scree plot	A plot of eigenvalues used to judge the number of factors to retain (the "elbow").
SD	Standard deviation.
SDQ	Strengths and Difficulties Questionnaire — a brief behavioural and emotional screen.
SE	Standard error — the standard deviation of a sampling distribution, indexing estimate precision.
Sensitivity	The proportion of true cases correctly identified by the measure.
SES	Socioeconomic status.
Specificity	The proportion of true non-cases correctly identified by the measure.
Split-half cross-validation	Randomly dividing a sample into two halves to derive a factor model on one half (EFA) and confirm it on the other (CFA); used here with the ROOTS cohort.
SR	Systematic review (also self-report in measurement contexts) — a structured synthesis of evidence using predefined, transparent methods.
SRMR	Standardised Root Mean Square Residual — a residual-based model-fit index (lower is better).
Strict / residual invariance	Scalar invariance plus equality of residual (error) variances across groups.
Structural equation modelling (SEM)	A framework for estimating relationships among latent and observed variables, encompassing CFA.
Structural validity	The degree to which scores reflect the hypothesised dimensional (factor) structure of the construct.
Subthreshold	Symptom patterns that are present but fall below the threshold required for a formal diagnosis.
Summed (sum) score	A total computed by adding item responses, commonly used in applied settings.
Tau-equivalence	The assumption that all items have equal factor loadings; the basis of Cronbach's alpha.
12-tier taxonomic audit	The structured framework developed in this portfolio to classify antisocial-behaviour measures from broadband externalising to narrow or specialised constructs.
Test–retest reliability	The stability of scores across repeated administrations over time.
Tetrachoric correlation	A correlation estimate appropriate for binary items, assuming an underlying continuous distribution.
TLI	Tucker–Lewis Index — an incremental model-fit index (≥ .95 indicates good fit).
Trait-based specifier	A dispositional qualifier (e.g., psychopathy) describing how/why behaviour occurs rather than the act itself; screened out by the taxonomic audit.
Transdiagnostic approach	An approach that cuts across diagnostic categories to capture variance shared by multiple disorders (e.g., the externalising framework).
TRF	Teacher's Report Form — teacher-report measure within the ASEBA family.
UEA	University of East Anglia.
Unidimensionality	The property that a set of items reflects a single underlying factor.

Validity	The extent to which a measure assesses the construct it intends to, and the appropriateness of inferences drawn from its scores.
Velicer's MAP test	Minimum Average Partial test — a method for deciding how many factors or components to retain.
VIF	Variance Inflation Factor — an index of multicollinearity among predictors.
W1/W2/W3	Wave 1 / Wave 2 / Wave 3 (data-collection timepoints).
WEIRD samples	Samples drawn from Western, Educated, Industrialised, Rich, and Democratic populations, which can limit global generalisability.
WLSMV	Weighted Least Squares Mean- and Variance-adjusted — an estimator suited to categorical/binary indicators.
"Worst score counts"	The COSMIN principle of rating a measurement property according to its weakest contributing element.
WRMR	Weighted Root Mean Square Residual — a residual-based fit index for categorical data (lower is better).
WSC	Worst Score Counts — the COSMIN algorithm that rates a measurement property by its weakest contributing element.
YLS/CMI	Youth Level of Service/Case Management Inventory.
Youden's index	A statistic identifying the threshold that best balances sensitivity and specificity.
YPI-S	Youth Psychopathic Traits Inventory – Short Version (a psychopathy measure noted in the review).
YSR	Youth Self-Report — youth-report measure within the ASEBA family.

Symbols and statistical notation

Symbol	Meaning
α	Cronbach's alpha — internal-consistency reliability assuming equal item loadings (tau-equivalence).
χ^2	Chi-squared — a model-fit/association statistic, sensitive to large sample sizes.
ΔCFI	Change in the Comparative Fit Index between nested models (used to judge measurement invariance).
ΔRMSEA	Change in RMSEA between nested models (used to judge measurement invariance).
I^2	I-squared — the proportion of between-study variability due to heterogeneity rather than chance.
κ	Cohen's kappa — chance-corrected inter-rater agreement.
ω	McDonald's omega — factor-model-based internal-consistency reliability (preferred when loadings vary).
r	Pearson correlation coefficient — strength of linear association between two continuous variables.
ρ	Spearman's rho — rank-order correlation for non-normal or ordinal data.

Chapter 1. Introduction to the Thesis Portfolio

This thesis portfolio presents research examining the measurement of adolescent antisocial behaviour (AB) and contemporary approaches to its conceptual and psychometric evaluation. The systematic review appraises existing instruments using two complementary frameworks: a structured 12-tier taxonomic audit developed to distinguish broadband externalising measures from narrow or specialised constructs, and alignment with the four behavioural domains of DSM-5 Conduct Disorder (CD) as a clinical reference framework. This sequential framework classifies instruments across 12 tiers by construct breadth, screening out trait-based specifiers (e.g., psychopathy), distal correlates (e.g., trauma), and clinical conditions (e.g., ADHD) so that only behavioural antisocial phenotypes are retained. Here, broadband externalising measures index the full spectrum of externalising behaviour (aggression, rule-breaking, oppositionality), whereas narrow or specialised constructs capture a single trait or context-specific domain (e.g., cyberbullying, dating violence). The subsequent empirical paper evaluates the psychometric properties of the Antisocial Behaviours Questionnaire (ABQ), a concise DSM-aligned measure, using factor analysis and measurement invariance testing across adolescence and emerging adulthood. Together, these studies aim to clarify methodological standards in the assessment of antisocial behaviour and to inform the selection of developmentally sensitive screening tools.

1.1 Antisocial Behaviour Across Development

Antisocial behaviour (AB) encompasses actions that violate social norms, legal rules, or the rights of others, ranging from rule-breaking and truancy to physical aggression and serious interpersonal harm (American Psychiatric Association [APA], 2013; Frick & Viding, 2009). These behaviours are of central concern within developmental

psychology, criminology, and educational psychiatry due to their association with adverse long-term outcomes, including substance misuse, educational disengagement, and later criminal involvement (Moffitt, 1993).

At the population level, antisocial behaviours follow the well-documented age–crime curve, characterised by an increase in mid-adolescence and a subsequent decline for many individuals (Moffitt, 1993). However, developmental research consistently identifies a subgroup of youth who demonstrate persistent trajectories associated with elevated long-term risk, with patterns that may extend from early childhood temperament through to adult offending (Caspi et al., 2014). Distinguishing normative, transient adolescent experimentation from early markers of persistent conduct pathology is therefore critical.

Importantly, adolescence does not represent a discrete endpoint in behavioural development. Emerging adulthood, now recognised as a prolonged transition to adult psychosocial functioning (Arnett, 2000; Sawyer et al., 2018), may involve shifts in the expression of antisocial behaviour—from overt school-based defiance to more covert acts such as theft or property damage, which carry greater legal and clinical implications. A developmentally informed framework must therefore accommodate both continuity and change across adolescence and young adulthood.

1.2 Antisocial Behaviour Within the Externalising Spectrum

Psychopathology is commonly organised into two broad higher-order dimensions: internalising and externalising difficulties. Internalising behaviours typically reflect inwardly directed symptoms such as anxiety and depression, whereas externalising behaviours involve outwardly directed actions characterised by behavioural disinhibition, rule-breaking, and aggression (Krueger et al., 2005; Lahey et al., 2012).

Within this broader externalising spectrum, antisocial behaviour represents a subset of actions that violate social norms, legal rules, or the rights of others.

Broadband externalising constructs often encompass a heterogeneous range of behaviours, including hyperactivity, impulsivity, oppositionality, and conduct-related problems. However, not all externalising behaviours are antisocial in nature. Antisocial behaviour is distinguished by its contextual and normative dimension—specifically, its transgression of socially established behavioural standards. This distinction becomes particularly salient in adolescence, where normative experimentation must be differentiated from patterns indicative of persistent conduct pathology.

Within this continuum, Conduct Disorder (CD) represents a categorical diagnostic threshold applied to minors. As defined in the DSM-5, CD requires the presence of at least three of fifteen behavioural symptoms across four domains: aggression to people or animals, destruction of property, deceitfulness or theft, and serious violations of rules (APA, 2013). CD therefore functions as a clinically defined point along the broader antisocial continuum, delineating behaviours of sufficient frequency and severity to warrant diagnosis.

Although dimensional models of externalising psychopathology (e.g., the Hierarchical Taxonomy of Psychopathology [HiTOP]; Kotov et al., 2017) conceptualise antisocial behaviours as manifestations of a broader latent liability, DSM-5 CD criteria provide a behaviourally explicit reference framework. The four-domain structure offers a clear organisational template for identifying the types of behaviours that constitute clinically significant conduct pathology. In the context of measurement, this framework can serve as a pragmatic guide for evaluating whether assessment

instruments comprehensively capture the breadth of antisocial behaviour relevant to adolescent populations.

Positioning antisocial behaviour within the broader externalising spectrum while recognising CD as a diagnostic threshold provides a conceptual foundation for examining how instruments operationalise construct coverage. It also clarifies the distinction between broadband externalising measures and those specifically targeting antisocial behaviours, a distinction that becomes critical when evaluating construct alignment and psychometric adequacy.

1.3 Challenges in Measuring Antisocial Behaviour

Despite longstanding recognition of antisocial behaviour as a critical developmental construct, its measurement remains fragmented. Four interrelated challenges characterise the current assessment landscape: instrument proliferation, tensions between dimensional and categorical measurement approaches, semantic inconsistency across measures, and the absence of a universally accepted benchmark for construct coverage.

First, the field contains a wide range of assessment tools, spanning ultra-brief screening instruments to comprehensive clinician-administered interviews (Frick et al., 2014). Brief measures prioritise feasibility and scalability, particularly in educational or community contexts, whereas longer instruments aim to support diagnostic formulation or forensic decision-making (APA, 2013). However, without a structured approach to evaluating content coverage and psychometric robustness, instrument selection may be influenced by accessibility, historical precedent, or administrative convenience rather than empirical precision (Farrington, 1995). This

proliferation complicates efforts to determine which tools adequately capture the breadth of antisocial behaviour.

Second, measurement approaches often reflect the broader conceptual tension between dimensional and categorical models. Some instruments are designed primarily to detect diagnostic thresholds, while others aim to index variation across the full behavioural continuum. Measures optimised for categorical identification may insufficiently represent subthreshold behaviours, whereas broadband dimensional scales may aggregate heterogeneous behaviours without ensuring comprehensive coverage of clinically relevant domains. This variation limits comparability across studies and settings.

Third, semantic inconsistency further obscures construct clarity. Identical subscale labels may refer to substantively different behavioural content across instruments, while different labels may capture overlapping behaviours (Streiner et al., 2015). For example, scales labelled “Conduct Problems” may emphasise irritability and defiance in one measure and serious rule violations in another. Reliance on subscale nomenclature alone therefore risks misinterpreting the behavioural constructs being assessed. Item-level scrutiny becomes necessary to determine whether instruments genuinely measure comparable behavioural phenomena.

Finally, the field lacks a definitive criterion against which antisocial behaviour measures can be evaluated (De Los Reyes, 2011). In the absence of a gold standard, construct alignment—the degree to which instrument items map onto theoretically defined behavioural domains—serves as a pragmatic proxy for validity (Haynes et al., 1995; Streiner et al., 2015). However, without a structured method for distinguishing broadband antisocial measures from narrow or context-specific instruments, evaluating such alignment becomes difficult. Together, these challenges underscore

the need for a systematic and transparent framework for assessing both content coverage and psychometric adequacy.

1.3 The Need for a Systematic Evaluation of Adolescent Antisocial Behaviour Measures

A substantial body of literature has examined antisocial behaviour and conduct problems in youth, including developmental trajectories, risk factors, comorbidity, and intervention efficacy (Frick & Viding, 2009; Moffitt, 1993). In parallel, evidence-based assessment frameworks have provided guidance on the clinical evaluation of disruptive behaviour disorders, outlining recommended instruments and multi-informant approaches (Mash & Hunsley, 2005). These contributions have been instrumental in advancing understanding of antisocial development and informing practice.

However, much of this literature has focused on conceptual models, treatment outcomes, or broader externalising behaviour problems rather than on the systematic evaluation of measurement instruments themselves. Where assessment guidance is provided, it often prioritises clinical utility or theoretical coherence over structured appraisal of measurement properties across independent studies. Similarly, existing reviews frequently aggregate externalising constructs (e.g., ADHD, oppositionality, conduct problems) without isolating antisocial behaviour as a distinct behavioural phenotype situated within developmental and normative contexts. This distinction matters because antisocial behaviour carries developmental, prognostic, and clinical implications that differ from adjacent externalising constructs: unlike ADHD (a regulatory/attentional construct) or oppositionality (defiance toward authority), antisocial behaviour specifically indexes the violation of others' rights and societal norms, which more directly predicts later offending and antisocial personality

outcomes (Moffitt, 1993; Frick & Viding, 2009). Aggregating these constructs therefore risks diluting measurement precision for a clinically high-stakes phenotype.

Furthermore, instrument-focused reviews that do exist are often not restricted to adolescence, span wide developmental age ranges, or focus on specific contexts (e.g., forensic settings or particular cultural adaptations). Given that adolescence represents the developmental period during which antisocial behaviour peaks for many individuals and differentiates into transient versus persistent trajectories (Moffitt, 1993; Caspi et al., 2014), a developmentally bounded evaluation of adolescent measures is warranted.

From a methodological perspective, reviews of youth mental health instruments across domains consistently demonstrate that internal consistency and some form of factor analysis are the most frequently reported properties, whereas cross-group comparability, measurement invariance, and comprehensive construct coverage are less consistently evaluated (Mokkink et al., 2018; Prinsen et al., 2018). In the context of antisocial behaviour, few syntheses have combined item-level content evaluation with structured psychometric appraisal frameworks such as COSMIN, nor have they integrated such appraisal with a clinically interpretable reference structure.

Crucially, the field lacks a systematic approach to determining whether instruments capture the breadth of behaviours constituting antisocial conduct in adolescence. In the absence of a definitive gold-standard benchmark (De Los Reyes, 2011), construct alignment—evaluating the extent to which instrument items map onto theoretically and clinically meaningful behavioural domains—becomes central to measurement validity (Haynes et al., 1995; Streiner et al., 2015). Yet prior reviews have not consistently operationalised construct coverage at the item level, nor distinguished

broadband antisocial measures from narrow or context-specific instruments before psychometric synthesis.

Taken together, these considerations indicate the need for a systematic, adolescent-specific evaluation of antisocial behaviour instruments that integrates (a) structured content filtering to isolate broadband measures, (b) domain-based construct alignment, and (c) rigorous psychometric appraisal across studies. The present portfolio addresses this gap through a multi-stage review process, followed by empirical validation of a concise measure aligned to the identified domain structure.

1.4 Evaluating Concise Measures Within Contemporary Psychometric Standards

The preceding discussion highlights that rigorous assessment of adolescent antisocial behaviour requires attention not only to psychometric performance, but also to construct coverage. As established earlier, antisocial behaviour sits within the broader externalising spectrum, yet not all externalising measures capture the full range of behaviours that constitute antisocial conduct. Some instruments focus on narrow expressions (e.g., specific contexts, traits, or isolated behavioural forms), whereas others aim to index the broader behavioural phenotype. Distinguishing between broadband and narrow operationalisations is therefore essential when evaluating the adequacy of any measure.

Importantly, breadth of coverage cannot be inferred from subscale labels or face validity alone. Instruments may share terminology while assessing substantively different behavioural content. Consequently, item-level scrutiny and structural evaluation are necessary to determine whether a measure meaningfully represents the behavioural domains it purports to assess. Without such examination, concise tools in

particular risk oversimplifying heterogeneous behaviours into superficially coherent but construct-limited indices.

At the same time, practical realities shape measurement choices in research and applied settings. Large-scale longitudinal cohorts and community-based studies often require brief self-report instruments that minimise respondent burden and permit repeated assessment across developmental stages. However, brevity introduces potential trade-offs. Reducing item count may constrain behavioural representation, alter dimensional structure, or limit comparability across gender and age if measurement invariance is not established.

These concerns are especially salient in adolescence and emerging adulthood, periods characterised by developmental shifts in behavioural expression. A measure that functions equivalently across age bands and between genders must demonstrate structural stability and invariance, rather than relying solely on internal consistency estimates. Prior use of an instrument in large samples does not, in itself, establish dimensional coherence, cross-group comparability, or diagnostic interpretability.

Accordingly, concise antisocial behaviour measures require systematic evaluation against contemporary psychometric standards. Such evaluation must extend beyond reliability to encompass dimensional structure, measurement invariance across developmental stages, construct validity, and clinically meaningful classification thresholds. Only through this comprehensive approach can it be determined whether a brief instrument achieves both practical feasibility and robust construct alignment.

1.5 Empirical Evaluation of the Antisocial Behaviour Questionnaire

The Antisocial Behaviour Questionnaire (ABQ) was originally developed as an 11-item self-report measure designed to capture concrete antisocial acts in adolescent

populations using a four-point Likert response format. The measure was implemented within two large UK community cohorts—the ROOTS study and the NeuroScience in Psychiatry Network (NSPN)—to index antisocial behaviours across adolescence and into emerging adulthood. Its concise format and behaviourally specific items positioned it as a practical tool for repeated assessment within longitudinal research designs.

Conceptually, the ABQ was intended to reflect behavioural domains consistent with conduct-related pathology, aligning with the broader construct considerations outlined earlier in this chapter. However, despite its application within substantial datasets, the measure has not previously undergone comprehensive psychometric evaluation using contemporary methodological standards. In particular, its dimensional structure, measurement invariance across gender and developmental stage, and clinically interpretable classification properties have not been systematically examined within a unified analytic framework.

As discussed in Section 1.4, prior use of a measure in large samples does not in itself establish structural coherence, cross-group comparability, or diagnostic utility. Concise instruments may offer clear feasibility advantages yet require careful evaluation to determine whether behavioural coverage and psychometric robustness are maintained when subjected to rigorous testing. This is particularly pertinent in adolescence, where developmental shifts in behavioural expression necessitate demonstration of stability across age and time.

Accordingly, the empirical component of this portfolio undertakes a comprehensive psychometric evaluation of the ABQ. This includes examination of dimensionality using exploratory and confirmatory factor analytic approaches, assessment of internal consistency, testing of measurement invariance across gender, age bands, and

longitudinal waves, evaluation of convergent and predictive validity, and analysis of diagnostic classification performance. Through this approach, the study assesses whether a concise antisocial behaviour measure can satisfy the construct alignment and methodological standards articulated in the preceding sections.

1.6 Aims of the Thesis Portfolio

This thesis portfolio has two primary aims.

The first aim was to conduct a systematic evaluation of instruments used to assess antisocial behaviour in adolescence. The review examined the extent to which existing measures demonstrate adequate construct coverage and psychometric robustness. This included structured content filtering to distinguish broadband antisocial measures from narrower constructs, alignment with clinically defined behavioural domains, and synthesis of evidence relating to structural validity, reliability, and cross-group comparability.

The second aim was to undertake a comprehensive psychometric evaluation of the Antisocial Behaviour Questionnaire (ABQ), a concise self-report instrument implemented in two large UK community cohorts. The empirical study examined the dimensional structure of the ABQ, internal consistency, measurement invariance across gender and developmental stage, construct and predictive validity, and diagnostic classification performance. The purpose was to determine whether a brief measure can meet contemporary standards of construct alignment and psychometric adequacy.

The portfolio is structured as follows. Chapter 2 presents the systematic review manuscript. Chapter 3 provides an extended methodological account and outlines the rationale for the empirical study. Chapter 4 presents the empirical manuscript

evaluating the ABQ. Chapter 5 provides additional methodological detail relating to the empirical analyses. Chapter 6 offers a final integrative discussion that synthesises findings from both components and considers implications for future research in adolescent antisocial behaviour assessment.

Chapter 2. A Systematic Review and Meta-Analysis of Adolescent Antisocial Behaviour Scales

Punita Panchal[1], Kenny Chiu[1], Sharon A S Neufeld[2]

[1]Department of Clinical Psychology and Psychological Therapies, Norwich Medical School, University of East Anglia, Norwich Research Park, Norwich NR4 7TJ, UK

[2]Department of Psychiatry, University of Cambridge, Cambridge CB2 8AH, UK

Abstract word count: 232

This manuscript has been formatted for submission to European Journal of Psychological Assessment (EJPA).

Manuscript word count based on EJPA guidelines (not including title page, acknowledgements & transparency page, and supplementary materials): 8853

Disclaimer/Declaration

The views and opinions expressed therein are those of the authors and do not necessarily reflect those of the NIHR, NHS or the Department of Health and Social Care.

Author Note

We have no conflicts of interest to disclose.

Correspondence concerning this article should be addressed to Sharon A S Neufeld, [2]Department of Psychiatry, University of Cambridge, Cambridge CB2 8AH, UK
Email: sasn2@medschl.cam.ac.uk

Abstract

Background: Precise measurement of adolescent antisocial behaviour is essential for effective assessment and intervention, yet the field is characterised by a diverse range of instruments of varying scope and methodological robustness. **Objectives:** This review aimed to identify instruments demonstrating comprehensive alignment with the four DSM-5 Conduct Disorder (CD) behavioural domains alongside adequate psychometric support, and to synthesise available evidence on structural validity, reliability, and measurement invariance. **Methods:** Following PRISMA and COSMIN guidelines, we conducted a systematic search across major databases and applied predefined inclusion and psychometric criteria. A structured 12-tier taxonomic audit was used to map item content to DSM-5 CD domains. Reliability generalisation meta-analysis was conducted for instrument families represented in ≥ 3 independent studies. **Results:** From 3,353 records, 275 studies representing 125 unique instrument variants were identified. Only 13 variants (10%) demonstrated both moderate-to-high psychometric certainty and full alignment with the four DSM-5 CD domains. Most measures relied primarily on Cronbach's alpha (79%) and principal component analysis (42%). Five instrument families met criteria for recommendation based on sufficient structural validity and internal consistency.

Conclusions: While internal consistency is frequently reported, comprehensive coverage of DSM-5 CD behavioural domains is less common. The identification of a small subset of DSM-aligned, psychometrically supported instruments provides structured guidance for researchers and clinicians selecting measures for diagnostic-level assessment and population-based screening of adolescent antisocial behaviour.

Keywords: antisocial behaviour; conduct problems; adolescents; psychometrics; systematic review.

Introduction

Assessment of adolescent antisocial behaviour (AB) is a cornerstone of developmental psychology, criminology, and educational psychiatry (e.g., Frick & Viding, 2009). These behaviours range from rule-breaking to physical aggression (American Psychiatric Association [APA], 2013) and predict long-term outcomes from substance abuse to incarceration (Moffitt, 1993). Although antisocial behaviours follow the well-documented age–crime curve, characterised by an increase in mid-adolescence and a subsequent decline for many individuals (Moffitt, 1993), a subgroup of youth demonstrate persistent trajectories associated with elevated long-term risk (Caspi et al., 2014), which can extend from early childhood temperament to adult offending. Precise measurement during the critical adolescent period is therefore essential for effective intervention. While this systematic review aims to identify scales that comprehensively measure antisocial behaviour, mapping this landscape highlights four significant challenges: (1) instrument proliferation, (2) the ongoing diagnostic versus dimensional debate, (3) the emergence of multiple nomenclatures, and (4) lack of a unified gold-standard benchmark. While these trends characterize a complex and maturing field, it can also collectively obscure the ability to assess the core antisocial phenotype. It is therefore vital to examine these trends to clarify what we measure, why we measure it and how to do so with the greatest clinical effectiveness.

Current assessments of the antisocial phenotype range from ultra-brief screens to exhaustive clinical interviews, each developed to meet distinct operational and setting requirements (Frick et al., 2014). While concise instruments prioritise administrative efficiency for rapid screening and large-scale triaging, more extensive measures are designed to provide the granularity necessary for establishing formal diagnoses or

informing complex forensic processes (e.g., Frick & Viding, 2009; APA, 2013). However, without a systematic method to filter these tools by content coverage and utility, selection is frequently driven by accessibility or historical precedent rather than empirical superiority (Farrington, 1995). Therefore, such proliferation necessitates a systematic audit to identify instruments that are both administratively efficient and psychometrically robust, ensuring they capture the comprehensive antisocial phenotype with the rigor necessary for clinical, forensic, and research applications.

Beyond the volume of instruments, the field faces a conceptual dichotomy between the dimensional construct of antisocial behaviour (AB) and the categorical diagnosis of Conduct Disorder (CD). Antisocial behaviour (AB) reflects a broad continuum of disinhibited behaviours, ranging from minor rule-breaking to severe aggression (Kotov et al., 2017). In contrast, CD is a psychiatric diagnosis for minors representing a specific “severity threshold” on this continuum, identified when a clinical line is drawn at a minimum of three symptoms alongside significant functional impairment (APA, 2013; Frick & Viding, 2009). This division is emblematic of a broader shift away from categorical syndromes, which are frequently hampered by high rates of clinical overlap and comorbidity (Kessler et al., 2005) and towards transdiagnostic approaches such as the externalising behaviour framework, which captures the common variance across diverse manifestations of psychopathology, exemplified by the 'p-factor' model (Caspi et al., 2014; Lahey et al., 2012; Neufeld, 2024). This dichotomy creates a hurdle for assessment: many current measures are optimized for categorical detection rather than dimensional mapping. Consequently, the field lacks a unified approach—one that offers spectrum-wide precision alongside incremental

diagnostic cut-offs to effectively guide support and intervention. Accordingly, this review will evaluate how effectively existing instruments navigate this continuum.

These conceptual divisions are further compounded by the same label being used for distinct constructs or different labels are used for the same behaviour. This makes it difficult to reliably determine if two instruments or their subscales with similar labels actually measure the same behaviour without looking at item-content or phrasing (Streiner et al., 2015). For example, 'Conduct Problems' subscale of the Strengths and Difficulties Questionnaire (SDQ) is composed of 5 items reflecting non-compliance and irritability (Goodman, 1997), whereas the same label in the Behavior Assessment System for Children-3 (BASC-3), a 139-item clinical tool, emphasizes serious rule-breaking and legal violations such as vandalism (Reynolds & Kamphaus, 2015). Consequently, a rigorous comparative synthesis must prioritise item-level alignment over canonical names to establish a "behavioural topography" that looks past the labels.

The final challenge is the absence of a universally accepted "gold-standard" benchmark for adolescent antisocial behaviour. Given the lack of a definitive criterion for AB (De Los Reyes, 2011), the most robust proxy for validity is construct alignment (Haynes et al., 1995; Streiner et al., 2015)—the degree to which an instrument's items comprehensively map onto theoretically established behavioural domains of antisocial behaviour, such as those mentioned in the DSM-5 Conduct Disorder criteria (APA, 2013). However, the field currently lacks a consensus-based taxonomy of these domains for antisocial behaviour, leaving researchers without standardized structural parameters for assessment. Consequently, alignment is often obscured by adjacent traits—like impulsivity or moral cognition—or instruments limited to specialized contexts like cyberbullying or dating violence. To isolate the

broadband externalising antisocial phenotype, this review filters out these peripheral components and maps the remaining instruments onto the four DSM-5 CD domains (aggression to people and animals, destruction of property, deceitfulness or theft, and serious violations of rules; APA, 2013). As the clinical anchor for the antisocial continuum, these four domains, comprising of 15 symptoms (APA, 2013) ranging from mild to severe, provide the structural framework to capture the subthreshold pattens often missed by categorical measures and identify exemplary, domain-anchored instruments that also demonstrate robust psychometrics.

The Present Review: Aim and Contribution

The present review conducts a comprehensive search of the literature to identify instruments with high construct alignment and robust psychometric standards. To achieve this we pursue two goals, (1) Content validity assessment to establish a "Behavioural Topography," this involves a two-stage filter: first, mapping instruments into a 12-tier taxonomic audit to isolate broadband externalising measures; and second, evaluating their structural alignment with the four DSM-5 CD symptom domains; (2) Psychometric Synthesis to evaluate factor structure, measurement invariance, and reliability of the selected instruments. For measures with ≥ 3 studies, a reliability generalisation was used to pool transformed internal consistency coefficients (Bonett, 2002) and synthesise structural validity evidence across diverse samples. Collectively, this ensures that the instruments used to identify at-risk youth capture the broad continuum of the antisocial phenotype while remaining anchored to clinical domains and possessing a level of psychometric precision that matches the complexity of their behaviour.

Methods

Systematic Review Protocol and Registration

This review was prospectively registered on PROSPERO (CRD420251039049) and conducted according to the PRISMA 2020 statement (Page et al., 2021) and COSMIN guidelines (Mokkink et al., 2018). Procedural transparency is detailed in the Open Science and Transparency statement.

The review transitioned through two technical phases:

(1) behavioural topography and content mapping, and

(2) psychometric synthesis and reliability generalisation.

Here, behavioural topography refers to item-level mapping of instrument content to predefined antisocial behavioural domains.

Information Sources and Search Strategy

A systematic search was conducted across MEDLINE (Ovid), PsycINFO (EBSCOhost), PubMed, and Scopus in November 2025, from database inception (no lower date limit was applied). Following COSMIN recommendations (Mokkink et al., 2018), the strategy utilized four conceptual blocks (Table S1): (1) Population (adolescents aged 12–18); (2) Construct (diagnostic and dimensional antisocial descriptors); (3) Instruments (evaluative scales); and (4) Psychometric Properties (Terwee et al., 2012).

To capture both subthreshold patterns and formal phenotypes (Frick et al., 2014; Kotov et al., 2017), search terms spanned clinical and developmental contexts using Medical Subject Headings (MeSH) and free text. No date limits were applied, and results were restricted to English-language peer-reviewed empirical studies reporting primary quantitative data on instrument development, validation, or translation. Full

syntax (Table S1) served as an anchor for all databases. Manual screening of reference lists supplemented the database search to ensure literature saturation. This identified *n* = 3,353 records.

Eligibility (Inclusion and Exclusion) Criteria

Eligibility was determined via a-priori criteria established per COSMIN guidance (Table S2; Mokkink et al., 2018). COSMIN (COnsensus-based Standards for the selection of health Measurement INstruments) is an internationally developed methodology and set of standards for evaluating the measurement properties (e.g., structural validity, internal consistency, and measurement invariance) of health-related and behavioural instruments (Mokkink et al., 2018). The target population was restricted to adolescents (approximately 12–18 years) across community, clinical, and forensic settings; mixed-age samples were included only if adolescent data were reported separately. Where disaggregated data were unavailable, inclusion was restricted to samples within a ±2-year developmental buffer (from ages 10–20 years); studies exceeding these boundaries (e.g., age 9–15 or 15–22 years) were excluded to maintain developmental homogeneity.

Regarding the construct, included instruments were required to assess primarily behavioural phenotypes (e.g., conduct, delinquency, aggression). To prevent construct drift, a 12-tier taxonomic audit (Table 1, S3) prioritized these phenotypes while excluding trait-based specifiers (e.g., psychopathy), distal correlates (e.g., trauma, ACEs), and clinical conditions (e.g., ADHD, ASD). Relational aggression and specific dyads (e.g., child-to-parent violence, dating violence) were retained for initial extraction but subjected to sequential taxonomic alignment to distinguish between broadband externalising behaviours and context-specific aggression (Table S4).

For instrument quality, studies were required to develop, translate, or validate a scale and report at least one psychometric property (e.g., internal consistency, structural validity, measurement invariance, test–retest reliability, construct validity, or criterion validity). Study design was limited to English-language, peer-reviewed empirical studies collecting primary data. Gray literature, including conference abstracts, unpublished theses, and non-empirical commentaries, was excluded to maintain a high evidence-level threshold (Table S5).

Quality and Psychometric Assessment

Methodological quality was appraised via the COSMIN Risk of Bias checklist (Mokkink et al., 2018). Following the "worst score counts" principle, each measurement property was assigned a four-point rating (Very Good, Adequate, Doubtful, or Inadequate).

Structural validity required confirmatory factor analysis (CFA) or equivalent latent modelling. Studies relying exclusively on Principal Component Analysis (PCA) were rated inadequate for structural validity, as PCA is a data-reduction technique that does not model latent constructs (Fabrigar et al., 1999).

Internal consistency was rated sufficient if Cronbach's $\alpha \geq .70$ or McDonald's $\omega \geq .70$, provided evidence of unidimensionality was present. Cronbach's alpha assumes equal factor loadings across items (tau-equivalence), which may not hold in multidimensional behavioural constructs.

Measurement invariance required at least metric-level support to demonstrate comparable factor loadings across groups (Putnick & Bornstein, 2016). Where items were analysed as categorical indicators, residual invariance was considered necessary for valid latent mean comparisons (Liu et al., 2017; Millsap & Yun-Tein, 2004).

Following this psychometric appraisal, the overall certainty of evidence was synthesised using the modified GRADE approach (Prinsen et al., 2018), which accounts for risk of bias, inconsistency, indirectness, and imprecision.

Statistical Synthesis and Reliability Generalisation

Quantitative synthesis was conducted following APA Meta-Analysis Reporting Standards (MARS; Appelbaum et al., 2018). Reliability generalisation was performed for instruments represented by ≥ 3 independent studies (k = 15; Bonett, 2010).

Reliability generalisation meta-analysis refers to the aggregation of internal consistency coefficients across independent samples to evaluate stability of reliability estimates across contexts.

Quantitative synthesis was conducted in R (Version 4.0.0; R Core Team, 2020) using the *metafor* package (Viechtbauer, 2010). To stabilise variance and account for the ceiling effect of reliability coefficients, raw alpha (α) values were transformed via the Bonett (2002) procedure prior to pooling. Pooled estimates and 95% confidence intervals (CIs) were derived using random-effects models, weighting each study by the inverse of its sampling variance.

Meta-analyses were stratified by informant type (e.g., Self, Parent, Teacher) to prevent conflation of distinct perspectives. Where internal consistency coefficients were reported for subscales rather than total scores, these were extracted and synthesised separately.

Statistical heterogeneity was quantified using Cochran's Q statistic and the I^2 metric (Higgins et al., 2003).

Results

Study Selection

The systematic search identified a total of 3,353 records. Following de-duplication (n = 975), 2,378 titles and abstracts were screened, resulting in the exclusion of 1,927 records. Of the 432 records assessed for eligibility against a priori criteria (Table S2), 157 were excluded (Table S4). Primary reasons for exclusion were administrative misalignment (n = 67; e.g., population age outside target range, non-empirical formats) and failure to satisfy the 12-tier taxonomic audit (n = 90; Table S3), which removed trait-based measures, distal correlates, and context-specific relational behaviours. A total of k = 275 studies were included (Figure 1). Because many studies evaluated multiple forms or subscales of the same measure, the analytic unit was disaggregated to n = 125 unique instrument variants (Study × Instrument units) for COSMIN evaluation (Mokkink et al., 2018). Following audits for (a) psychometric properties (Table 3) and (b) alignment with the four DSM-5 CD domains (APA, 2013), 13 instrument variants met criteria for final synthesis (Table 4). These 13 variants comprised adolescent self-report, caregiver-report, and clinician/teacher-report forms distributed across five instrument families.

Description of Studies and Instruments

Characteristics of Included Studies: Across k = 275 studies (309 Study × Instrument units), sample sizes ranged from N = 50 to 297,102. The evidence base represented 22 countries, primarily clustered in North America, Europe, and East Asia (Table 1). Most evaluations utilised mixed-gender samples. Pooled samples (N = 542,430) comprised 54.2% males; however, male representation rose to 82% within forensic cohorts. Targeted populations included school-based (n = 84), forensic (n = 26), and

clinical (n = 21) samples. While all focused on adolescence (12–18 years), longitudinal designs captured developmental trajectories from late childhood (Moffitt, 1993). Advanced latent modelling approaches (e.g., confirmatory and higher-order models) were more frequently observed in more recent publications (Table S8).

Characteristics of Measurement Instruments: A total of n = 125 unique instrument variants were identified (Table 2). Measurement was dominated by adolescent self-report (n = 102), supplemented by parent (n = 18), teacher (n = 6), or clinician (n = 5) perspectives. Scales ranged in length from 5 items to 186 items. Major construct domains included general aggression, psychopathic traits (callous-unemotional, narcissism, impulsivity), and overt delinquency (Frick et al., 2014; Table S3). Cultural adaptations comprised a substantial proportion of the pool, notably in Portuguese, Chinese, and Arabic (Table S8).

Methodological Quality of the Evidence Base

Methodological quality was audited using the COSMIN Risk of Bias checklist and the "Worst Score Counts" (WSC) algorithm (Table S7; Mokkink et al., 2018). This stage evaluated the procedural rigor of the n = 125 instrument variants, with findings for each measurement property consolidated in the full evidence synthesis (Table S8) and summarised visually in the distribution of risk of bias ratings (Figure 2).

Structural Validity and Internal Consistency: Internal consistency achieved the highest proportion of "Very Good" ratings (89%, n = 117), with coefficients typically exceeding the ≥ .80 threshold (Table 3). However, 79% of evaluations relied on Cronbach's alpha, with only 6% reporting McDonald's omega (Table S8). Both Cronbach's alpha and McDonald's omega are coefficients of internal consistency reliability (the degree to which the items of a scale measure the same underlying

construct); alpha assumes that all items contribute equally to the construct (tau-equivalence), whereas omega is estimated from a fitted factor model and is generally preferred when this assumption does not hold (McNeish, 2018). Structural validity was a primary source of methodological limitation. Forty-two percent (n = 55) of evaluations were rated "Inadequate" due to reliance on Principal Component Analysis (PCA). As PCA is a data-reduction technique and does not model latent constructs, it does not provide evidence of structural validity (Fabrigar et al., 1999). A substantial proportion of CFA-based studies did not report all conventional fit indices; 48% did not report SRMR (or WRMR where applicable; SRMR = standardised root mean square residual, WRMR = weighted root mean square residual) and 41% did not report TLI. Thirty-six percent (n = 47) of CFA-based studies did not satisfy conventional model fit thresholds (e.g., CFI < .90). Model fit statistics were extracted as reported; the extent to which post-hoc model modifications (e.g., correlated residuals or cross-loadings) were applied was not systematically coded in the present synthesis. Only a small subset of instruments tested higher-order or bifactor structures, which were rated "Very Good" under COSMIN criteria based on reported fit indices. It should be noted that bifactor models are known to often yield improved fit indices relative to correlated-factor models, and the present review did not evaluate comparative model performance.

Cross-Cultural Validity and Measurement Invariance: Measurement invariance (MI) evidence was reported in 32% (n = 42) of instrument variants. Metric invariance was most commonly reported. Scalar invariance was less frequently reported. Measurement invariance levels were coded based on reported model comparisons; estimator type (e.g., maximum likelihood [ML] versus weighted least squares mean- and variance-adjusted [WLSMV]) was not systematically extracted. Absence of scalar

or residual invariance evidence reflects lack of reported support rather than confirmed model failure. Cultural adaptations were frequently rated "Doubtful" due to limited reporting of translation procedures and absence of reported multi-group CFA to establish cross-language equivalence.

Reliability: Quality for test–retest reliability was frequently capped at "Doubtful" (Table 3). This was primarily due to retest intervals exceeding four weeks. Clinician-rated instruments achieved higher inter-rater reliability ratings (ICC ≥ .80). In contrast, inter-rater reliability was rarely evaluated in self-report measures.

Construct and Criterion Validity: Construct validity evidence was generally strong, with 74% of evaluations meeting a priori hypotheses (Table 3). Criterion validity was frequently rated as Adequate where instruments were evaluated against established self-report measures with clinically derived cut-offs; however, independent gold standards (e.g., structured diagnostic interviews or judicial records) were less commonly used. Under strict COSMIN criteria, absence of an independent gold standard limits the strength of criterion validity evidence; ratings here reflect pragmatic benchmarking practices within the field.

Synthesis of Results: Quality of the Instruments

Of the *n* = 125 unique variants extracted, 45 (34%) satisfied the predefined psychometric floor, achieving moderate-to-high GRADE quality for structural validity and internal consistency. Full alignment with all four DSM-5 CD behavioural domains was then applied as an additional criterion. This second filter reduced the pool to 13 variants meeting both psychometric sufficiency and comprehensive domain coverage. These 13 variants were clustered within five broader instrument families (e.g., ASEBA, Proposed Specifiers for Conduct Disorder [PSCD]) (Table 4).

Reliability Generalisation: A reliability generalisation meta-analysis was performed for k = 15 instrument families (Table 5). The pooled estimate for Cronbach's alpha across antisocial behaviour constructs was α = .81 (95% CI [.79, .83]), indicating sufficient aggregate internal consistency. Substantial heterogeneity was observed for most self-report measures (I^2 > 90%; values above 75% are typically considered substantial; Higgins et al., 2003), justifying GRADE downgrading for consistency. In contrast, clinician-rated instruments demonstrated comparatively lower heterogeneity.

Summary of Measurement Invariance: Measurement invariance (MI) data represented the most significant evidence gap in the synthesis. While 34% (n = 45) of variants provided some reported evidence of invariance, the quality of evidence was frequently downgraded under GRADE criteria due to limited reporting of scalar invariance (Table 3). A small cluster (e.g., SBPS, BCS-A, Ungdata Scale; Bakken et al., 2016) provided high quality evidence for scalar invariance, permitting valid mean comparisons across gender and age. Most accepted instruments (e.g., RPQ, Inventory of Callous-Unemotional Traits; ICU-11; Frick, 2004; PSCD) demonstrated sufficient metric invariance, confirming that the underlying antisocial constructs are interpreted consistently across groups; however, mean comparisons remain indeterminate without reported scalar evidence.

Measurement invariance levels were coded based on reported model comparisons; estimator type (e.g., ML versus WLSMV) was not systematically extracted. Absence of scalar or residual invariance evidence reflects lack of reported support rather than confirmed model failure.

Notably, several cultural adaptations (e.g., Buss-Perry Aggression Questionnaire-Short Form; BPAQ-SF Czech version; Bryant & Smith, 2001; Šmira & Šerek, 2021) achieved moderate quality evidence for age and gender invariance but lacked reported

cross-national equivalence data (Table S8), limiting their utility for international comparative research.

Hypothesis Testing for Construct and Criterion Validity: The majority of synthesised instruments (74%) established sufficient evidence (+) for construct validity by meeting ≥ 75% of a priori hypotheses (Table 3). Convergent validity was robustly supported with high quality evidence for instruments measuring callous-unemotional traits (e.g., ICU-11, Youth Psychopathic Traits Inventory-Short Version; YPI-S; van Baardewijk et al., 2010) and aggression (e.g., RPQ), with correlations consistently exceeding the r = .50 threshold. Criterion validity was frequently rated as Adequate where instruments were evaluated against established self-report measures with clinically derived cut-offs (e.g., using the YSR to validate the YEBS); however, independent gold standards (e.g., clinician-led structured interviews or judicial records) were less commonly used (Table S8). Under strict COSMIN criteria, absence of an independent gold standard limits the strength of criterion validity evidence; ratings here reflect pragmatic benchmarking practices within the field. Significant informant discrepancies in construct and criterion validity were identified; parent-report structures (e.g., BASC PRS-A) exhibited lower risk of bias and higher GRADE certainty than youth self-reports when validated against clinical diagnoses (Table 4).

Feasibility and Practical Utility: Feasibility significantly influenced the final COSMIN categorisation (Table 2) and clinical recommendation. Clinician-rated tools (e.g., PCL-YV, YLS/CMI) provided high-quality evidence for reliability but were constrained by high administrative burden, requiring specialised training and extended administration time. In contrast, brief self-report screeners (e.g., ICU-11, SBPS) were prioritised for scalability, requiring less than 10 minutes for completion with no training costs (Table 4). The highest practical utility for universal screening was

identified in tools with free accessibility and extensive multi-language translations (e.g., YSR, SDQ, PSCD), facilitating cross-cultural research without licensing barriers. However, the synthesis noted a feasibility gap: tools with higher GRADE certainty for structural validity often required longer administration times, suggesting a trade-off between psychometric rigour and routine clinical feasibility (Table 4).

Final Recommendations and Categorization

Instruments were categorised using COSMIN guidelines (Mokkink et al., 2018) into Category A (Recommended for use; sufficient evidence for structural validity and internal consistency), Category B (Potential for use; requiring further validation), and Category C (Not recommended; insufficient evidence for key properties). The final classification was determined by a dual-criterion audit of psychometric certainty and behavioural congruence (Table 3). Category A ratings reflect psychometric sufficiency under COSMIN criteria; however, final recommendation additionally required full alignment with all four DSM-5 CD behavioural domains. Of the n = 125 unique variants, 34% (n = 45) satisfied the psychometric floor, achieving moderate-to-high GRADE quality for structural validity, internal consistency, and reported measurement invariance. Behavioural alignment requiring coverage across all four DSM-5 CD domains (APA, 2013) reduced this pool to 13 variants clustered within five instrument families (Table 4). Among these, the PSCD and the Achenbach System of Empirically Based Assessment (ASEBA; Achenbach, 1991; consisting of CBCL/YSR/TRF systems) represent the most psychometrically robust multi-informant options. Notably, while the APA CD symptom list satisfied the alignment criterion, it yielded low GRADE quality evidence, indicating a need for more standardised validation despite its high content validity (Table 4).

Discussion

Principal Findings

This systematic review and reliability generalisation meta-analysis provides a comprehensive audit of adolescent antisocial behaviour instruments filtered through a sequential 12-tier taxonomic audit and a COSMIN-aligned psychometric lens (Mokkink et al., 2018; Tinghög et al., 2024). Although numerous instruments purport to measure antisocial behaviour or related constructs, applying structured taxonomic mapping, DSM-5 Conduct Disorder (CD) domain alignment, and psychometric quality criteria resulted in substantial reduction in eligible measures. From an initial 3,353 records, the 12-tier taxonomic audit narrowed the evidence base to k = 275 studies representing 125 unique instrument variants (Figure 1). Subsequent application of dual criteria—moderate-to-high GRADE psychometric certainty and coverage across all four DSM-5 CD domains (APA, 2013)—further reduced this pool to 13 variants clustered within five instrument families, representing 10% of the original evidence base. This pattern indicates that while many tools are available and frequently used, relatively few demonstrate both structural robustness and comprehensive behavioural coverage. The reliability generalisation meta-analyses (Table 5; Figure 3) yielded pooled $\alpha = .81$ (95% CI [.79, .83]) across 15 instrument families, indicating adequate aggregate internal consistency. However, substantial heterogeneity ($I^2 > 90\%$ for most self-report measures) suggests that reliability varies meaningfully across contexts and samples. Although 89% of evaluations exceeded the $\alpha \geq .80$ threshold, alignment with the full behavioural topography of DSM-5 CD was less frequently achieved. Among the 13 variants meeting both criteria, the PSCD and the Achenbach System of Empirically Based Assessment (ASEBA) represent the most psychometrically robust multi-informant instrument families. Collectively, these findings indicate that while instruments measuring antisocial behaviour are plentiful, only a small subset demonstrate both comprehensive domain coverage and sufficient psychometric

support. By prioritising content alignment and multi-informant integration, this review clarifies which instrument families are best positioned to support consistent assessment in clinical practice and research settings.

Quality of the Measurement Instruments

The synthesis indicates that administrative brevity often coincides with reduced coverage of the full behavioural range specified within DSM-5 Conduct Disorder (CD). Although many instruments are widely used to assess antisocial behaviour, structured DSM-5 CD content mapping demonstrated that numerous variants lack sufficient breadth to capture all four diagnostic domains. Adequate assessment across adolescence requires sufficient item spread and density to reflect behaviours ranging from lower-frequency rule-breaking to more serious and persistent offending (APA, 2013; Rijlaarsdam et al., 2024). Brief screeners (e.g., SBPS, ICU-11) prioritise efficiency but may omit domains necessary for comprehensive construct representation. This trade-off is reflected in the reliability generalisation findings (Table 5), where longer broadband instruments (e.g., YSR, CBCL) demonstrated pooled reliability of $\alpha \approx .91$, whereas shorter screeners such as the SDQ yielded lower pooled internal consistency ($\alpha \approx .58$). Interpretation of these differences must also consider informant perspective. Caregiver-report instruments (e.g., BASC PRS-A, ASEBA) generally achieved higher GRADE certainty and lower risk of bias than youth self-reports; however, this does not imply inherent superiority. Antisocial behaviours frequently occur in peer, community, or unsupervised contexts that may not be directly observed by caregivers or teachers. In such cases, adolescent self-report captures behavioural manifestations that would otherwise remain unmeasured, albeit with greater statistical heterogeneity (De Los Reyes et al., 2015). A further distinction emerged between instruments developed using contemporary latent modelling approaches and those relying on earlier analytic strategies. Measures supported by confirmatory factor analysis (CFA), bifactor modelling, or exploratory structural equation

modelling (ESEM) (e.g., PSCD, CPAQ-R) demonstrated clearer differentiation between general antisociality and more specific behavioural dimensions (Morin et al., 2020). However, bifactor and higher-order models are known to produce comparatively favourable global fit indices relative to simpler structures (Greiff & Heene, 2017), and most applied settings rely on summed subscale scores rather than model-estimated factor scores. As such, structural robustness demonstrated at the latent level does not automatically translate to optimal performance under sum-score usage, particularly when tau-equivalence assumptions are not met (McNeish, 2018). Importantly, these patterns also reflect historical shifts in psychometric practice. Earlier validation studies commonly relied on principal component analysis and Cronbach's alpha as primary indicators of structure and reliability. Over the past two decades, increased computational capacity and methodological development have facilitated broader adoption of CFA, multi-group invariance testing, and alternative reliability estimators such as McDonald's omega (Hayes & Coutts, 2020; Putnick & Bornstein, 2016). Consequently, variability in reported structural rigor may partly reflect evolving analytic standards rather than intrinsic instrument limitations. Finally, the synthesis highlights a continuing tension between psychometric depth and practical feasibility. Clinician-administered instruments (e.g., PCL-YV) demonstrate strong reliability but require substantial training and administration time (Table 2). In contrast, brief screeners maximise feasibility but may compromise domain coverage. Taken together, the findings support multi-informant batteries in which structurally robust caregiver reports and context-sensitive adolescent self-reports are integrated to provide comprehensive behavioural assessment across settings.

Methodological Quality of the Evidence Base

Beyond individual instrument properties, the synthesis identified considerable variability in methodological approaches across the antisocial behaviour measurement literature. Although

confirmatory modelling techniques have been available for several decades (Brown, 2015), 42% of structural evaluations relied on principal component analysis (PCA), a method intended for data reduction rather than latent construct modelling (Fabrigar et al., 1999). This pattern likely reflects historical analytic conventions, particularly in earlier validation studies conducted prior to the widespread adoption of confirmatory factor analysis within applied psychology.

Similarly, 79% of studies relied exclusively on Cronbach's alpha as an indicator of internal consistency. While alpha remains widely reported, its assumptions of tau-equivalence and sensitivity to scale length have led to increasing recommendations for complementary indices such as McDonald's omega (Dunn et al., 2014; Hayes & Coutts, 2020). The limited reporting of alternative reliability estimators therefore appears consistent with broader historical reporting norms rather than methodological oversight. Measurement invariance was evaluated in approximately one-third of instrument variants. Where reported, metric invariance was more commonly supported than scalar invariance. The absence of scalar or residual invariance evidence in many studies constrains confidence in mean comparisons across demographic groups; however, multi-group CFA procedures and routine invariance testing have only more recently become standard components of validation practice (Putnick & Bornstein, 2016). Variability in invariance reporting thus likely reflects evolving methodological expectations over time. Reporting of model fit indices also varied. Among CFA-based studies, 48% did not report SRMR (or WRMR when categorical estimators were used), and 41% did not report TLI, limiting direct cross-study comparability. Contemporary psychometric guidelines increasingly recommend reporting a comprehensive set of fit indices to facilitate transparent evaluation. Taken together, these findings suggest that validation practices within antisocial behaviour research span multiple methodological periods. Continued integration of multi-group CFA, alternative reliability estimators, and

comprehensive fit reporting would likely enhance interpretability and comparability in future instrument development. These developments can be understood as part of the broader evolution of psychometric standards rather than deficiencies specific to this field.

Strengths and Limitations

A major strength of this review lies in its structured integration of behavioural content mapping with psychometric appraisal. By combining COSMIN-aligned evaluations (Table 3) with systematic mapping against the four DSM-5 Conduct Disorder (CD) domains (Table S3), the synthesis identifies instruments demonstrating both methodological robustness and comprehensive behavioural coverage. The 12-tier taxonomic audit enhanced conceptual precision by distinguishing core antisocial behaviours from distal correlates. Our meta-analyses encompassed 125 unique instrument variants (Table 2) representing over 540,000 participants, and generated pooled reliability estimates across 15 instrument families (Table 5; Figure 3), providing both quantitative synthesis and a structured overview of the measurement landscape. Several limitations should be considered. First, DSM-5 CD alignment represents a deliberate comparative framework. Although it enhances diagnostic coherence, alternative dimensional models of antisocial behaviour may prioritise overlapping but distinct constructs. Similarly, despite predefined criteria, taxonomic classification required interpretative judgement at the item level, particularly for instruments positioned near conceptual boundaries. While structured, this process cannot be entirely free from coder influence. Second, the evidence base is geographically skewed toward WEIRD populations, and the restriction to English-language, peer-reviewed studies may limit global representativeness. Third, the application of relatively stable psychometric benchmarks to adolescent behaviour presents conceptual challenges. COSMIN retest criteria assume stability across short intervals; however, behavioural frequency and expression may fluctuate during adolescence due to developmental and contextual factors (Trull et al., 2021). Lower test–

retest coefficients may therefore reflect behavioural change rather than measurement inadequacy. Finally, COSMIN and GRADE ratings are dependent on the level of methodological detail reported in primary studies, and instrument categorisation reflects established psychometric thresholds. Where properties were not tested or not fully reported, ratings were necessarily conservative. Conclusions regarding criterion validity are further constrained by the frequent reliance on self-report proxies rather than independent external standards. Finally, the included studies were conducted across diverse national and cultural contexts (the evidence base represented 22 countries; Table 1); cross-cultural differences in the expression, interpretation, and reporting of antisocial behaviour may have contributed to the substantial between-study heterogeneity observed and should be considered when interpreting the pooled estimates.

Together, these considerations contextualise the findings while supporting the review's central contribution: a transparent framework for evaluating adolescent antisocial behaviour instruments according to both behavioural coverage and psychometric support.

Conclusions

This review addressed a field characterised by a diverse array of instruments of varying scope and methodological robustness assessing adolescent antisocial behaviour. By anchoring comparison to the four DSM-5 Conduct Disorder (CD) domains and integrating structured psychometric evaluation, the synthesis provides a transparent framework for appraising measures in relation to both diagnostic content coverage and empirical support. Rather than resolving broader conceptual debates, this approach clarifies which instruments demonstrate comprehensive alignment with DSM-5 CD behavioural domains alongside sufficient psychometric evidence. This distinction is particularly relevant when differentiating between brief screening tools and broadband diagnostic instruments. The practical implications are direct. At the population level, clearer differentiation between screeners and DSM-aligned

instruments may support more appropriate screening strategies. In clinical and forensic settings, use of measures demonstrating multi-domain coverage and structural stability may enhance consistency in assessment and interpretation across developmental stages and demographic groups. Overall, the review contributes a reproducible, DSM-anchored reference framework to support more coherent instrument selection and improved comparability across research and applied contexts.

References

Achenbach, T. M. (2017). Future directions for clinical research, services, and training: Lessons from ASEBA outcomes. *Journal of Clinical Child & Adolescent Psychology, 46*(1), 139–145. https://doi.org/10.1080/15374416.2016.1212363

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin, 101*(2), 213–232. https://doi.org/10.1037/0033-2909.101.2.213

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

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). https://doi.org/10.1176/appi.books.9780890425596

Andershed, H., Kerr, M., Stattin, H., & Levander, S. (2002). Psychopathic traits in non-referred youths: A new assessment tool. In E. Blaauw & L. Sheridan (Eds.), *Psychopaths: Current international perspectives* (pp. 125–158). Elsevier.

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist, 73*(1), 3–25. https://doi.org/10.1037/amp0000191

Archer, R. P., Handel, R. W., Ben-Porath, Y. S., & Tellegen, A. (2016). *Minnesota Multiphasic Personality Inventory–Adolescent–Restructured Form (MMPI-A-RF): Manual for administration, scoring, and interpretation*. University of Minnesota Press.

Bonett, D. G. (2002). Sample size requirements for estimating Cronbach's alpha. *Journal of Educational and Behavioral Statistics, 27*(4), 335–340. https://doi.org/10.3102/10769986027004335

Bonett, D. G. (2010). Varying coefficient meta-analytic methods for alpha reliability. *Psychological Methods, 15*(4), 368–385. https://doi.org/10.1037/a0020142

Buss, A. H., & Perry, M. (1992). The aggression questionnaire. *Journal of Personality and Social Psychology, 63*(3), 452–459. https://doi.org/10.1037/0022-3514.63.3.452

Calvete, E., Orue, I., & Sampedro, R. (2013). Development and validation of the Child-to-Parent Aggression Questionnaire. *Journal of Family Violence, 28*(7), 693–705. https://doi.org/10.1007/s10896-013-9528-7

Canivez, G. L., Reynolds, C. R., & Kamphaus, R. W. (2021). Construct validity of the BASC-3 Teacher Rating Scale: Hierarchical exploratory factor analyses of the child and adolescent forms. *Journal of Psychoeducational Assessment, 39*(3), 312–330. https://doi.org/10.1177/0734282920974801

Caspi, A., Houts, R. M., Belsky, D. W., Goldman-Mellor, S. J., Harrington, H., Israel, S., Meier, M. H., Ramrakha, S., Shalev, I., Poulton, R., & Moffitt, T. E. (2014). The p factor: One general psychopathology factor in the Dunedin Multidisciplinary Health and Development Study? *Clinical Psychological Science, 2*(2), 119–137. https://doi.org/10.1177/2167702613497473

Caspi, A., Moffitt, T. E., & colleagues. (2024). Developmental trajectories of psychopathology across the life course. *Annual Review of Clinical Psychology, 20*, 1–25.

Chai, K., et al. (2021). Efficiency of AI-prioritization in systematic reviews. *Journal of Information Science*. https://doi.org/10.1177/0165551521102551

Colins, O. F., Vermeiren, R. R., Schuyten, G., & Fanti, K. A. (2014). Psychometric properties of the Inventory of Callous-Unemotional Traits in detained female adolescents. *Psychological Assessment, 26*(2), 509–520. https://doi.org/10.1037/a0035515

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions*. University of Illinois Press.

De Los Reyes, A. (2011). Introduction to the special section: More than measurement error: Discovering meaning in informant discrepancies. *Journal of Clinical Child & Adolescent Psychology, 40*(1), 1–9. https://doi.org/10.1080/15374416.2011.533405

De Los Reyes, A., Augenstein, T. M., Wang, M., Thomas, S. A., Drabick, D. A., Burgers, D. E., & Rabinowitz, J. (2015). The eighteen-year update: Outcomes of professional efforts to improve understanding of the informant discrepancy phenomenon. *Journal of Clinical Child & Adolescent Psychology, 44*(1), 1–15. https://doi.org/10.1080/15374416.2014.930680

De Los Reyes, A., & Kazdin, A. E. (2005). Informant discrepancies in the assessment of childhood psychopathology: A critical review, theoretical framework, and recommendations for further study. *Psychological Bulletin, 125*(4), 483–509. https://doi.org/10.1037/0033-2909.125.4.483

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ, 315*(7109), 629–634. https://doi.org/10.1136/bmj.315.7109.629

Estévez, E., Jiménez, T. I., & Moreno, D. (2013). Aggressive behavior in adolescence as a predictor of child-to-parent violence. *Journal of Adolescence, 36*(4), 673–684. https://doi.org/10.1016/j.adolescence.2013.05.002

Farrell, A. D., Kung, E. M., White, K. S., & Valois, R. F. (2000). The structure of self-reported aggression, drug use, and delinquent behaviors during early adolescence. *Journal of Clinical Child Psychology, 29*(2), 282–292. https://doi.org/10.1207/S15374424jccp2902_13

Farrington, D. P. (1995). The development of offending and antisocial behaviour from childhood. *Journal of Child Psychology and Psychiatry, 36*(6), 929–964. https://doi.org/10.1111/j.1469-7610.1995.tb01342.x

Forth, A. E., Kosson, D. S., & Hare, R. D. (2003). *The Psychopathy Checklist: Youth Version (PCL:YV).* Multi-Health Systems.

Frick, P. J. (2012). Developmental pathways to conduct disorder. *Journal of Clinical Child & Adolescent Psychology, 41*(3), 378–389. https://doi.org/10.1080/15374416.2012.664815

Frick, P. J., & Hare, R. D. (2001). *Antisocial Process Screening Device: APSD.* Multi-Health Systems.

Frick, P. J., Ray, J. V., Thornton, L. C., & Kahn, R. E. (2014). Callous-unemotional traits and serious conduct problems. *Psychological Bulletin, 140*(1), 1–57. https://doi.org/10.1037/a0033076

Frick, P. J., & Viding, E. (2009). Antisocial behavior from a developmental psychopathology perspective. *Development and Psychopathology, 21*(4), 1111–1125. https://doi.org/10.1017/S095457940999002X

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

Goodman, R. (2001). Psychometric properties of the Strengths and Difficulties Questionnaire. *Journal of the American Academy of Child & Adolescent Psychiatry, 40*(11), 1337–1345. https://doi.org/10.1097/00004583-200111000-00015

Greiff, S., & Heene, M. (2017). Why psychological assessment needs to start worrying about model fit. *European Journal of Psychological Assessment, 33*(5), 313–317. https://doi.org/10.1027/1015-5759/a000450

Grisso, T., & Barnum, R. (2001). *Massachusetts Youth Screening Instrument–Second Version (MAYSI-2): User's manual and technical report.* University of Massachusetts Medical School.

Hakstian, A. R., & Whalen, T. E. (1976). A K-sample significance test for independent alpha coefficients. *Psychometrika, 41*(2), 219–231. https://doi.org/10.1007/BF02294150

Hayes, A. F., & Coutts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. *Communication Methods and Measures, 14*(1), 1–24. https://doi.org/10.1080/19312458.2020.1718629

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ, 327*(7414), 557–560. https://doi.org/10.1136/bmj.327.7414.557

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis. *Structural Equation Modeling, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Jesness, C. F., & Wedge, R. F. (1970). *The Jesness Inventory–Revised.* Multi-Health Systems.

Kelley, T. L. (1927). *Interpretation of educational measurements.* World Book Company.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310

Marsee, M. A., Barry, C. T., Childs, K. K., Frick, P. J., Kimonis, E. R., Muñoz, L. C., Aucoin, K. J., & Lau, K. S. (2011). Assessing aggression using self-report: Factor structure and invariance of the Peer Conflict Scale. *Psychological Assessment, 23*(3), 792–804. https://doi.org/10.1037/a0023369

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23*(3), 412–433. https://doi.org/10.1037/met0000144

Moffitt, T. E. (1993). Adolescence-limited and life-course-persistent antisocial behavior. *Psychological Review, 100*(4), 674–701. https://doi.org/10.1037/0033-295X.100.4.674

Morin, A. J. S., Arens, A. K., & Marsh, H. W. (2020). A bifactor exploratory structural equation modeling framework. *Structural Equation Modeling, 27*(3), 397–417. https://doi.org/10.1080/10705511.2019.1688449

Nielsen, M., Haun, D., Kärtner, J., & Legare, C. H. (2017). Persistent sampling bias in developmental psychology. *Journal of Experimental Child Psychology, 162*, 31–38. https://doi.org/10.1016/j.jecp.2017.04.017

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

Page, M. J., et al. (2021). The PRISMA 2020 statement. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71

Patrick, C. J. (2010). Operationalizing the triarchic conceptualization of psychopathy. Florida State University.

Pechorro, P., Marôco, J., Ray, J. V., & Gonçalves, R. A. (2015). Psychometric properties of the Social Conduct Scale. *International Journal of Offender Therapy and Comparative Criminology, 59*(12), 1283–1301. https://doi.org/10.1177/0306624X14546768

Pascual-Sagastizabal, E., Azurmendi, A., Sánchez-Martín, J. R., & Braza, F. (2019). Psychometric properties of the revised Child-to-Parent Aggression Questionnaire (CPAQ-R). *European Journal of Psychological Assessment, 35*(3), 338–347. https://doi.org/10.1027/1015-5759/a000399

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting. *Developmental Review, 41*, 71–90. https://doi.org/10.1016/j.dr.2016.06.004

Reynolds, C. R., & Kamphaus, R. W. (2015). *Behavior Assessment System for Children* (3rd ed.). Pearson.

Shaffer, D., Fisher, P., Lucas, C. P., Dulcan, M. K., & Schwab-Stone, M. (2024). The Adolescent Functioning Scale: Development and validation. *Journal of Child Psychology and Psychiatry*. Advance online publication.

Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment, 12*(1), 102–111. https://doi.org/10.1037/1040-3590.12.1.102

Streiner, D. L., Norman, G. R., & Cairney, J. (2015). *Health measurement scales* (5th ed.). Oxford University Press.

Terwee, C. B., et al. (2007). Quality criteria for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology, 60*(1), 34–42. https://doi.org/10.1016/j.jclinepi.2006.03.012

Thomas, H. J. (2019). Measuring bullying in adolescence. *Journal of Adolescence, 74*, 1–12.

Tinghög, P., Malm, A., Arwidson, C., Sigvardsdotter, E., Lundin, A., & Saboonchi, F. (2024). Prevalence of exposure to potentially traumatic events and PTSD in the general population. *European Journal of Psychotraumatology, 15*(1), Article 2331994. https://doi.org/10.1080/20008066.2024.2331994

Trull, T. J., Lane, S. P., & Carpenter, R. W. (2021). Developmental instability and personality pathology. *Annual Review of Clinical Psychology, 17*, 355–381. https://doi.org/10.1146/annurev-clinpsy-081219-103304

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

Vreugdenhil, C., et al. (2006). Psychometric properties of the Dutch version of the DISC-IV. *European Child & Adolescent Psychiatry, 15*(1), 25–35. https://doi.org/10.1007/s00787-006-0501-1

1. Acknowledgements

The authors would like to express their sincere gratitude to Kellsey McCann for their assistance with the initial title and abstract screening phase of this systematic review. SN was supported by the Wellcome Trust Early Career Award (226392/Z/22/Z). All research in the Department of Psychiatry at Cambridge University benefits from the National Institute for Health Research (NIHR) Cambridge Biomedical Research Centre. Finally, we acknowledge the use of Rayyan for machine-assisted screening support.

2. Funding

This work was supported by the University of East Anglia, Norwich, UK. No specific grant from any funding agency in the public, commercial, or not-for-profit sectors was received for this research.

3. Conflicts of Interest

The authors declare that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

4. Open Science Protocol and Registration:

This review was prospectively registered on PROSPERO (ID: CRD420251039049). *Data Transparency:* The raw psychometric data extraction sheet and exclusion log are provided as Electronic Supplementary Material (ESM). *Materials Transparency:* The 45-variable extraction framework and coding definitions are available in the ESM. *Analytic Code Transparency:* All R-code (metafor package) used for the reliability generalisation is provided in the ESM. *Transparency Statement:* This review adheres to PRISMA-S and COSMIN guidelines.

Table 1.

Landscape of Assessment Instruments and Sample Characteristics Grouped by Taxonomic Category

Category	Included Instruments (Variants/Subscales)	*k*	Total *N* (Range)	Age *M*(*SD*) Range	Primary Population	Items	Informants	Languages Covered
A1. Broad Measures	CBCL (Standard, DSM), YSR (Standard, DSM), SDQ, SDQ-DP, SDQ-IS, BASC-PRS (A/C), BASC-SRP, BASC-BESS, AFS, SCPS, SCS, YI-4, YIEPS, YEPS, S-SBQ, TSR	114	142,500 (102–26,306)	10.7 (1.4)–17.2 (1.1)	Community / Clinical	10–138	S, P, T, C	EN, ES, PT, DE, FR, ZH, UR, IT, IS, MT, NL, NO
B1. Conduct / Del.	DISBA, STAB, SRD, ASRDS, AHSRD, BPC, CD-CL, ECBI, ISRD Items, PBFS, OADP-CDS, RBPC, UNGDATA Scale, ABQ, BQAB	29	12,450 (138–5,083)	12.0 (1.2)–17.2 (0.9)	Community / Risk	6–89	S, P, T, C	EN, ES, PT, NO, DE
C1. Gen. Aggression	AQ (29-item), AQ-SF, BAQ, BWAQ, BWAQ-ML, APBFS, BDHI, EBS, LHA, OMS, R-MOAS	24	8,900 (73–3,996)	12.4 (1.5)–17.6 (1.2)	Community / Forensic	9–75	S, P, T	EN, ES, PT, ZH, NL, EL, IT, AR
D1. React./Proact.	RPQ, RPQ-P, PCS, PCS-20, C-SHARP, IRPA, AAS, IPAS, Little (2003) measure	24	15,200 (101–5,615)	11.0 (1.6)–16.7 (1.3)	Forensic / School	12–52	S, P, T	EN, DE, FR, ZH, MS, IT, PT, HU
E1. Bullying / Cyber	BCS-A, ECIPQ, ECIP-Q, EBIP-Q, FBS (V/P), OBVQ-R, SBPS/SBVS, BullyHARM, CBQ, CBVEQ-G, CGQ-A, HAT, RPEQ, ZBBS	18	11,800 (397–3,044)	10.6 (1.6)–15.9 (1.2)	Community (School)	10–64	S, Peer	EN, ES, ZH, EL, PT, IT, DE
F1. Child-to-Parent	CPV-Q, CPAQ, CPAQ-R, RAQ	12	6,500 (169–2,719)	12.0 (2.0)–16.2 (1.6)	Risk / Community	18–33	S, P	ES, ZH, EN
G1. CU Traits / Psy.	ICU (Standard, -11, -12, -10), APSD (SR, PR), PCL:YV, CPS, YPI, YPI-S, YPI-SV, INCA, INCA-M, CAPE 1.1, MACI, MTI, PAS	79	24,600 (67–9,797)	10.6 (3.2)–17.7 (1.7)	Forensic / Clinical	4–66	S, P, T, C	EN, PT, ZH, DE, ES, NL, FA, FI, HE
G2. Triarchic Psy.	TriPM, Tri-Scales, MMPI-A-RF (Tri-Scales)	3	3,250 (660–1,540)	13.0 (1.8)–15.7 (1.1)	Community / Court	18–58	S, P	EN, IT
H1. Forensic Risk	YLS/CMI, MAYSI-2, FER-R, PACT, AssetPlus, SAVRY, OC	10	12,800 (154–3,906)	12.0 (1.8)–17.2 (1.4)	Forensic (Adjudicated)	24–53	S, C, Staff	EN, ES, PT, NL

Note. *k* = number of independent studies; *N* = total participants; M(SD) = mean age and standard deviation (where reported) in years. Categories A1–H1 follow the *a priori* taxonomic framework designed to isolate direct behavioural antisocial phenotypes from mechanisms or contextual correlates. Informant codes: S = Self-report, P = Parent/Caregiver, T = Teacher, Peer = Peer-nomination, C = Clinician (file review/interview), Staff = Institutional staff observation. Language Codes (ISO 639-1): EN (English), ES (Spanish), PT (Portuguese), ZH (Chinese), DE (German), FR (French), IT (Italian), NL (Dutch), NO (Norwegian), EL (Greek), UR (Urdu), AR (Arabic), MS (Malay), IS (Icelandic), MT (Maltese), FA (Persian), FI (Finnish), HE (Hebrew), HU (Hungarian).

Table 2.

Directory and Structural Characteristics of Included Measurement Instruments

Acronym	Instrument Full Name	Developer Citation	Items	Informants	Age Range	k
ASEBA	Achenbach System (CBCL, YSR, TRF)	Achenbach & Rescorla (2001)	103–118	S, P, T, C	10–21	24
ICU	Inventory of Callous-Unemotional Traits	Frick (2004)	10–24	S, P, T	10–20	21
SDQ	Strengths & Difficulties Questionnaire	Goodman (1997)	25	S, P, T	10–19	16
YPI	Youth Psychopathic Traits Inventory	Andershed et al. (2002)	18–50	S	12–19	13
RPQ	Reactive-Proactive Aggression Questionnaire	Raine et al. (2006)	23	S, P	11–20	12
AQ	Aggression Questionnaire (incl. BPAQ/BWAQ)	Buss & Perry (1992)	12–29	S	11–20	10
PSCD	Proposed Specifiers for Conduct Disorder	Frick (2012)	12–29	S, P	10–19	10
APSD	Antisocial Process Screening Device	Frick & Hare (2001)	20	S, P, T	10–21	8
PCL:YV	Psychopathy Checklist: Youth Version	Forth et al. (2003)	20	C	12–21	7
CPAQ/CPV	Child-to-Parent Aggression/Violence Qs	Calvete / Estévez (2013)	20–28	S, P	12–18	7
Bullying	BCS-A, ECIP-Q, EBIP-Q, ZBBS	Various (e.g. Thomas, 2019)	10–64	S, Peer	10–18	15
BASC	Behavior Assessment System for Children	Reynolds & Kamphaus (2015)	27–160	S, P, T	11–18	4
YLS/CMI	Youth Level of Service/Case Management	Hoge & Andrews (2002)	42	C	12–18	4
PCS	Peer Conflict Scale	Marsee et al. (2011)	20–40	S, P	11–17	3
STAB	Subtypes of Antisocial Behavior Q	Burt & Donnellan (2009)	32	S	11–18	2
PBFS	Problem Behavior Frequency Scale	Farrell et al. (2000)	15–35	S	11–17	2
MAYSI-2	Massachusetts Youth Screening Instrument	Grisso & Barnum (2001)	52	S	12–17	2
MESSY	Matson Eval. of Social Skills with Youngsters	Matson et al. (1983)	62	S, P, T	12–17	1
TriPM	Triarchic Psychopathy Measure	Patrick (2010)	58	S	13–18	1
MIA	Mental Health & Social Inadaptation Assessment	Côté et al. (2017)	113	S	15	1
INCA	Inv. of CU Traits and Antisocial Behavior	Nas et al. (2012)	28–33	S	13–19	1
SCS	Social Conduct Scale	Pechorro et al. (2015)	80	S	11–17	1
AFS	Adolescent Functioning Scale	Shaffer et al. (2024)	33	P	11–18	1

Note. k = number of independent studies identifying the measurement engine. Age Range reflects the inclusive span across all extracted reports (incorporating the ±2y developmental buffer). Informants: S: Self-report; P: Parent/Caregiver; T: Teacher; C: Clinician/Record review. ASEBA (Achenbach System of Empirically Based Assessment) includes CBCL, YSR, TRF, and YSR-AB variants. AQ (Aggression Questionnaire) family includes BPAQ and BWAQ adaptations. CPAQ/CPV represents specialized measures for child-to-parent aggression. Bullying category pools specialized instruments for victimization and perpetration (e.g., ECIP-Q, BCS-A). Structural features represent the most frequently reported item-count configuration.

Table 3.

COSMIN Synthesis: Methodological Quality and Evidence Certainty by Instrument Family

Acronym	Structural Validity	Internal Consistency	Stability / Reliability	Construct Validity	Criterion Validity	GRADE Quality	COSMIN Category	CD Alignment
ASEBA	Very Good	Very Good	Adequate	Very Good	Very Good	High	A	Strong
PCL:YV	Adequate	Very Good	Very Good	Very Good	Very Good	High	A	Strong
PSCD	Very Good	Very Good	Adequate	Very Good	NR	High	A	Strong
BASC	Adequate	Very Good	Adequate	Doubtful	Very Good	Moderate	B	Strong
MIA	Very Good	Very Good	Inadequate	Very Good	NR	Moderate	B	Strong
YPI	Very Good	Very Good	Adequate	Very Good	Very Good	High	A	Partial
STAB	Very Good	Adequate	Adequate	Very Good	NR	High	A	Partial
YLS/CMI	Inadequate	Adequate	Very Good	Very Good	Very Good	High	B	Partial
RPQ	Very Good	Very Good	Inadequate	Very Good	Very Good	High	A	Partial
SDQ	Adequate	Adequate	Adequate	Adequate	Inadequate	Moderate	B	Partial
CPAQ/CPV	Very Good	Adequate	Inadequate	Very Good	Adequate	Moderate	B	Partial
PCS	Adequate	Adequate	NR	Adequate	NR	Moderate	B	Partial
INCA	Very Good	Very Good	Inadequate	Very Good	NR	Moderate	B	Partial
MESSY	Adequate	Very Good	Inadequate	Very Good	NR	Moderate	B	Partial
PBFS	Very Good	Inadequate	Very Good	Very Good	NR	Moderate	C	Partial
APSD	Inadequate	Adequate	Inadequate	Adequate	Doubtful	Low	C	Partial
MAYSI-2	Inadequate	Adequate	Inadequate	Adequate	Doubtful	Low	C	Partial
SCS	Inadequate	Very Good	Inadequate	Very Good	NR	Low	C	Partial
AFS	Inadequate	Very Good	Adequate	Very Good	NR	Low	C	Partial
TriPM	Very Good	Adequate	Adequate	Very Good	NR	High	A	None
ICU	Adequate	Very Good	Adequate	Very Good	Doubtful	Moderate	B	None
AQ	Very Good	Very Good	Adequate	Very Good	Adequate	Moderate	B	None
Bullying	Very Good	Very Good	Inadequate	Adequate	Doubtful	Moderate	B	None

Note. Quality ratings (Very Good to Inadequate) follow COSMIN "worst score counts" logic. GRADE Quality indicates confidence in the evidence. COSMIN Category: A: Recommended (valid/reliable); B: Potential (requires data); C: Not Recommended (poor properties). Rated against Table S5: Structural (CFI/TLI ≥ .90, RMSEA ≤ .08), Consistency (α\ω ≥ .70 with unidimensionality), and Reliability (ICC ≥ .70). NR: Not Reported.

Table 4.

Operational Directory of High-Alignment (Five Instrument Families) Diagnostic Instruments

Instrument Family	Variants Represented	Age Range	Informants	Primary Reference	Subscales (Full List)	Languages & Geographic Variants	Admin Time	Training	Cost
ASEBA	CBCL, YSR, TRF, CBC	10–21	S, P, T, C	Achenbach & Rescorla (2001)	Syndromes: Aggressive, Rule-Breaking, Anxious/Depressed, Withdrawn, Somatic, Social, Thought, Attention. DSM: Conduct, ODD, Affective, Anxiety, ADHD.	English, Spanish, German, Chinese, Portuguese, Finnish, Dutch, French, Arabic, Urdu, Malay.	15–20m	Low	Paid
PSCD	PSCD-P, PSCD-SV, PSCD-Y	10–19	S, P	Frick (2012)	Dimensions: CU Traits (Uncaring, Callousness, Unemotional), Narcissism (Grandiose-Manipulative), Impulsivity (Conduct Problems).	English, Persian, Dutch, Mandarin, Portuguese, Spanish, French.	10m	Low	Free
PCL:YV	PCL-YV (18 & 20 item)	12–21	C	Forth, Kosson, & Hare (2003)	Factor 1: Glibness, Grandiosity, Lying, Manipulative, No Remorse, Shallow Affect, No Empathy, No Responsibility. Factor 2: Seeking Stimulation, Parasitic, No Goals, Impulsive, Irresponsible, Poor Control, Early Behavior, Delinquency.	English, Spanish, Portuguese, French, Dutch, Swedish.	90–120m	High	Paid
MIA	MIA (Self-report)	12–17	S	Côté et al. (2017)	Symptoms: CD, ODD, ADHD, GAD, MDD. Behavior Clusters: Aggression, Theft, Rule-Breaking, Social Inadaptation.	French, English (Primarily validated in Quebec/Canada).	25m	Low	Free
BASC	PRS, SRP, TRS, BESS	11–18	S, P, T	Reynolds & Kamphaus (2015)	Clinical: Aggression, Conduct, Hyperactivity, Anxiety, Depression, Somatic, Atypicality, Withdrawal, Attention. Adaptive: Adaptability, Social Skills, Leadership, Study, Functional Comm.	English, Spanish (Broadly used in USA/Europe).	20–30m	Mod	Paid

Note. Informants: S: Self-report; P: Parent; T: Teacher; C: Clinician/Professional. Training: Low: Manual-based administration; Mod: Requires graduate-level training in psychometrics; High: Requires formal certification and interview expertise (PCL:YV). Languages: Lists the primary validated translations identified within the systematic search.

Table 5.

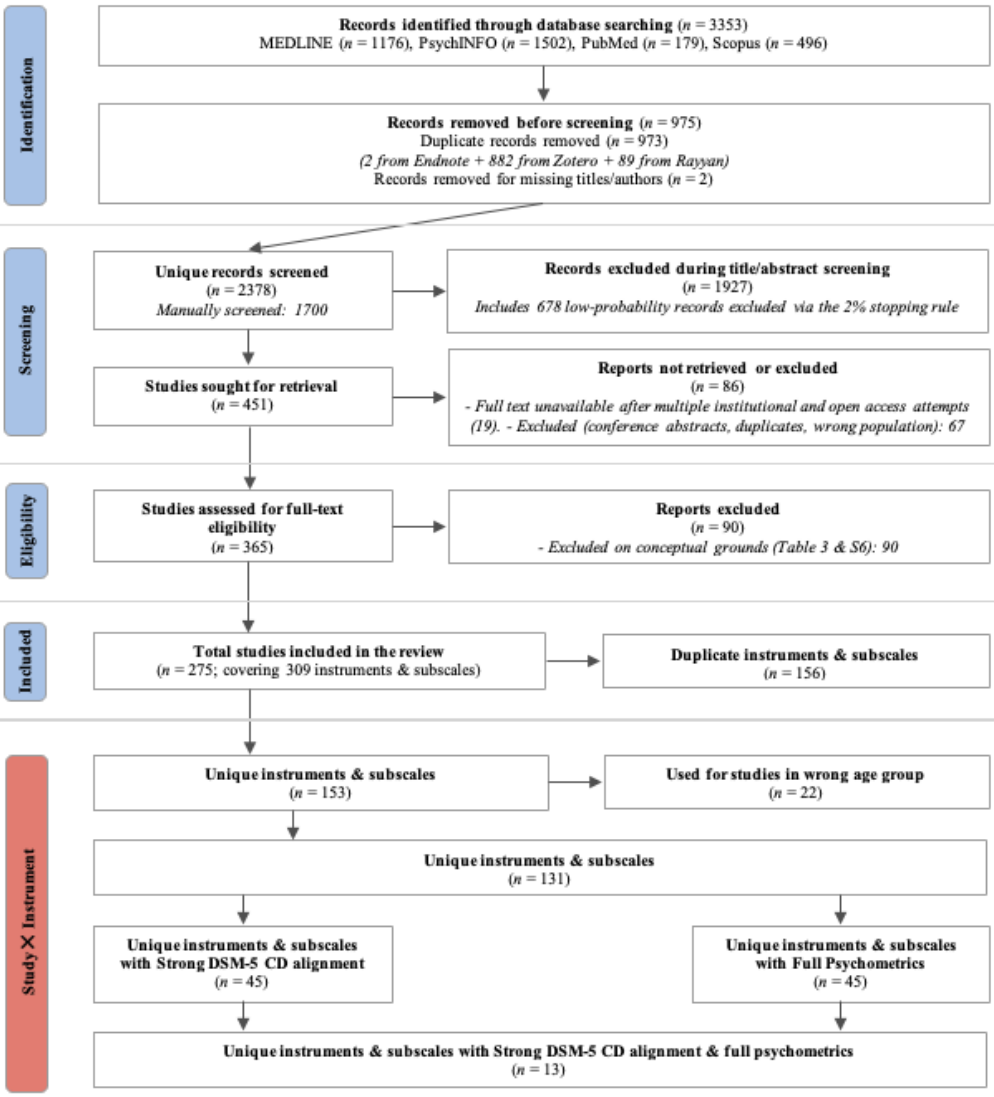
Reliability Generalisation Meta-Analysis Results Categorised by Taxonomic Alignment

Instrument Family	Alignment	*k*	Total N	Pooled α	95% CI	I^2 (%)
Strong CD Alignment						
CBCL	Strong	3	1,037	.91	[.87, .94]	94.1
PSCD	Strong	7	6,103	.90	[.88, .92]	95.7
PCL-YV	Strong	3	530	.88	[.86, .90]	53.2
YSR	Strong	8	7,198	.91	[.88, .94]	98.7
Partial CD Alignment						
YPI	Partial	12	5,777	.89	[.87, .91]	95.2
YPI-S	Partial	10	7,018	.81	[.79, .82]	79.2
ICU	Partial	16	13,967	.81	[.80, .83]	92.2
RPQ	Partial	11	17,225	.88	[.86, .89]	96.1
APSD	Partial	8	5,909	.76	[.67, .82]	98.0
APSD-SR	Partial	7	7,255	.67	[.59, .74]	96.4
SDQ	Partial	10	43,948	.58	[.49, .66]	99.5
SDQ-S	Partial	4	17,886	.58	[.49, .65]	98.7
Supplementary Scales						
AQ-SF	—	3	1,640	.88	[.80, .93]	97.8
AQ	—	3	2,397	.84	[.62, .93]	99.5
CPV-Q	—	3	2,538	.84	[.53, .95]	99.7

Note. k = number of independent samples; Total N = aggregate participant count; pooled α = pooled estimate of Cronbach's alpha derived from random-effects meta-analysis following Bonett (2002) transformation; CI = Confidence Interval; I^2 = statistical heterogeneity quantifying the proportion of variance attributable to study-level differences.

Figure 1.

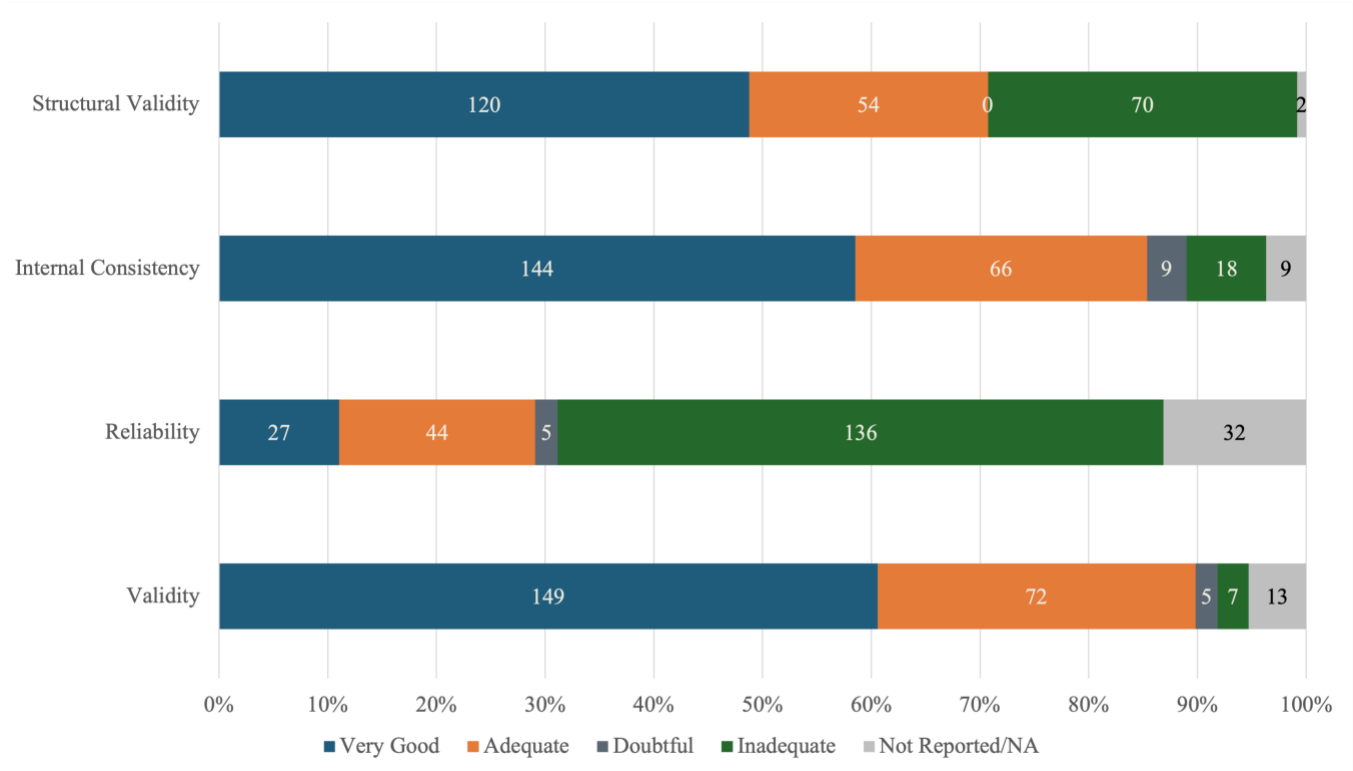
PRISMA 2020 Flow Diagram of the Systematic Search and Selection Process for Adolescent Antisocial Behaviour Instruments



Note. Diagram illustrates the identification and filtering of records. Of the full-text exclusions, $n = 90$ were removed via a 12-tier taxonomic audit to isolate broadband antisocial phenotypes. The unit of analysis shifted from records ($k = 275$) to unique instrument variants ($n = 125$) for psychometric appraisal. Final synthesis focused on $n = 13$ variants (representing 5 instrument families) satisfying the dual-criterion filter of psychometric sufficiency and full DSM-5 Conduct Disorder alignment.

Figure 2.

Distribution of Methodological Quality (Risk of Bias) Ratings for Measurement Properties across 125 Instrument Variants



Note. Ratings are assigned based on the COSMIN "worst score counts" principle. Bars represent the proportion of instrument variants achieving specific quality levels for each reported property. The high prevalence of "Inadequate" ratings for Structural Validity reflects the widespread reliance on Principal Component Analysis (PCA), while "Doubtful" ratings for Invariance and Reliability highlight the systematic omission of scalar testing and appropriate retest intervals.

Supplementary Materials

Table S1.

Electronic Database Search Strategy: Concept Blocks, Controlled Vocabulary, and Free-Text Terms

Concept Block		Controlled Vocabulary (MeSH / Publication Type)	Key Free-Text Terms (title/abstract)
I.	Behaviour	Conduct Disorder/; Juvenile Delinquency/	antisocial behavio?r, anti-social behavio?r, conduct problem*, rule-break*, delinquen*, externaliz*, externalis*, aggression, aggressive behavio?r
II.	Instrument	Surveys and Questionnaires/; Psychological Tests/; Interview, Psychological/; Self Report/	questionnaire*, survey*, scale*, rating scale*, instrument*, measure*, assessment tool*, screening tool*, checklist*, inventory*, index*, subscale*, self-report*, structured interview*, semi-structured interview*, schedule*, battery*
III.	Psychometrics	Psychometrics/; Reproducibility of Results/; Validation Study.pt.	reliab*, valid*, psychometric*, internal consistency, Cronbach*, omega, ICC, kappa, test–retest, factor analys*, CFA, EFA, measurement invariance, IRT, Rasch, DIF, cross-cultural, translation, adaptation
IV.	Population	Adolescent/	adolescen*, teen*, youth, young person*, young people

Note. Syntax reflects Ovid (MEDLINE/PsycINFO) format; similar platform-specific syntax was used across other databases. Controlled vocabulary refers to MeSH adapted for specific platforms. Free-text terms searched in titles/abstracts using wildcards (*) and proximity operators. Blocks combined as (I AND II AND III AND IV); terms within blocks combined using OR. Search conducted November 2025 from inception. Limits: English language and peer-reviewed only.

Table S2.

Eligibility Criteria for Study Selection and Operational Decision Codes

Category	Inclusion Criteria	Exclusion Criteria	Relevant Screening Labels
Population (Age & Setting)	Adolescents (≈ 12–18 years). Mixed samples acceptable if adolescent subgroup clear or predominant. Settings: community, clinical (CAMHS), forensic/justice, or social-care (foster, residential).	Exclusively adult or child (< 10 yrs) samples with no separate adolescent data. University or adult offender samples. Purely educational performance settings unlinked to behavioural/clinical aims.	Include – Psychometric Study (Antisocial Behaviour) Exclude – Wrong Population (Age) Exclude – Wrong Population (Setting)
Construct (What is measured)	Instruments assessing antisocial, conduct, delinquent, aggressive, rule-breaking, bullying, or externalising behaviour. May include peer violence, sexual harassment, child-to-parent violence if behaviourally focused.	Instruments measuring unrelated constructs (depression, anxiety, wellbeing, optimism, resilience, lifestyle, trauma, ACEs, autism, microaggressions). Instruments focused on personality traits/disorders (e.g. ASPD, psychopathy, PID-based).	Include – Psychometric Study (Antisocial Behaviour) Exclude – Not Antisocial Behaviour (Unrelated Construct) Exclude – Not Antisocial Behaviour (Personality Focus)
Instrument & Psychometrics	Studies that develop, adapt, translate, or validate an antisocial-behaviour instrument. Report ≥ 1 psychometric property (e.g., reliability, validity, factor analysis, invariance, Rasch/IRT, sensitivity/specificity).	Studies using a scale only to describe, correlate, or predict outcomes without reporting psychometric properties. No reliability/validity or structural analysis. Teacher/clinical observations without standardised measures.	Exclude – Not Psychometric Exclude – Observational / Non-Psychometric Ratings
Study Design / Aim	Quantitative studies of instrument development or validation in adolescents.	Interventions, case studies, or correlational/longitudinal risk-factor analyses (e.g., SES → delinquency) without psychometric evaluation. Theoretical papers, reviews, qualitative-only studies.	Exclude – Intervention Only Exclude – Theoretical / Review / Commentary Exclude – Qualitative Only
Language & Publication Type	English-language, peer-reviewed empirical papers providing sufficient methodological detail to extract psychometric data.	Non-English papers; conference abstracts; unpublished theses; retracted or withdrawn publications.	Exclude – Not English Language Exclude – Conference Abstract / No Full Text Retracted / Withdrawn Study
Unclear / Mixed Cases	Abstracts mentioning psychometric terms but lacking detail; mixed age or unclear construct—retain for full-text review.	–	Unclear – Insufficient Info Unclear – Mixed Sample Unclear – Possibly Related Construct
Administrative	–	Duplicates or secondary citations.	Duplicate Reference / Review for Hand search

Note. Eligibility criteria were established *a priori.* Screening labels correspond to the operational decisions applied during title and abstract screening. Studies that did not clearly meet criteria based on abstract information were marked as “Unclear” and retained for full-text "Personality Focus" excludes instruments measuring underlying traits (e.g., psychopathy) without a specific behavioural phenotype.

Table S3.

Taxonomy of Antisocial Construct Categories and Inclusion Rationale

Category	Subcategory	Include / Exclude	Description / Rationale	*k*
A. General / Broad Behaviour Measures	A1. Core Broad Instruments	INCLUDE	Broad externalising/behaviour problem screeners (e.g., CBCL, YSR, SDQ, BASC) with validated antisocial/aggression subscales.	81
	A2. Psychometric / General Risk Refinements	EXCLUDE	General psychopathology or personality scales not specific to antisocial behaviour.	8
B. General Antisocial / Conduct Measures	B1. Conduct / Delinquency Scales	INCLUDE	Direct measures of conduct problems, delinquency, and antisocial acts.	29
	B2. Core Underlying Constructs	EXCLUDE	Measures of mechanisms (e.g., impulsivity, EF) rather than antisocial behaviour itself.	8
C. Aggression (General)	C1. General Aggression Scales	INCLUDE	Direct measures of aggressive behaviour relevant to antisocial constructs.	24
	C2. Anger / Hostility / Irritability	EXCLUDE	Affective/internalising states unless validated against conduct outcomes.	7
D. Reactive vs Proactive Aggression	D1. Reactive / Proactive Instruments	INCLUDE	Functional subtypes of aggression essential for risk assessment.	24
E. Bullying & Cyberbullying	E1. Bullying / Cyberbullying Measures	INCLUDE	Peer aggression and rule-breaking in school/online contexts.	18
	E2. Dating / Sexual / Partner Aggression	EXCLUDE	Specialised aggression domain outside general antisocial behaviour.	8
F. Child-to-Parent Violence	F1. CPV Instruments	EXCLUDE	Measures of youth-perpetrated aggression toward caregivers.	7
G. CU Traits & Youth Psychopathy	G1. Psychopathy / Callous–Unemotional Traits	INCLUDE	Core predictors of severe, persistent antisocial behaviour.	79
	G2. Triarchic Model Measures	INCLUDE	Modern psychopathy frameworks (TriPM).	3
H. Justice-Involved Risk Assessment	H1. Risk / Recidivism Tools	INCLUDE	Forensic tools predicting antisocial behaviour and reoffending.	10
I. Substance Use / Risk Behaviour	I1. Substance Use / Risk-Taking Scales	EXCLUDE	Measures alcohol/drug use or general risk-taking, not antisocial behaviour.	12
	I2. Clinical / Treatment-Context Tools	EXCLUDE	Instruments tracking treatment processes rather than antisocial constructs.	6
	J1. Empathy / Moral Cognition	EXCLUDE	Mechanisms underlying antisocial behaviour, not behavioural outcomes.	7

Category	Subcategory	Include / Exclude	Description / Rationale	*k*
J. Empathy, Moral Cognition, Prosocial Functioning	J2. Social / Prosocial Functioning	EXCLUDE	Social competence measures unless primary validation focuses on antisocial behaviour.	4
K. Parenting / Family / Social Climate	K1. Family / Parenting Measures	EXCLUDE	Environmental/contextual risk measures (e.g., parenting style).	7
	K2. Peer / Social Context Measures	EXCLUDE	Peer relations and social climate predictors, not antisocial behaviour measures.	5
L. Special Populations (ADHD/ASD/ID)	L1. Developmental / Clinical Condition Scales	EXCLUDE	Measures of developmental conditions unless containing antisocial subscales.	8
	L2. Specific Symptom Measures	EXCLUDE	Narrow behavioural symptoms (e.g., ODD, hyperactivity) not capturing the antisocial construct.	11

Note. *k* = Number of studies. Table S4 presents the hierarchical taxonomy used to classify all 365 records assessed at the full-text stage. Categories and subcategories distinguish measures that directly assess antisocial behaviour or its recognised subdomains (included) from constructs considered correlates, precursors, or contextual influences (excluded). A total of 275 studies met inclusion criteria across the identified taxonomic clusters.

Table S4.

List of Studies Excluded at the Full-Text Assessment Stage with Rationales

Author (Year)	Exclusion Stage	Primary Rationale for Rejection	Specific Tier/Reason
Caspi et al. (2001)	Eligibility	Instrument not main focus	Evaluates informant agreement on fathers' behaviour; not instrument validation.
Connolly et al. (2017)	Eligibility	Instrument not main focus	Focus on latent class analysis rather than psychometric properties of a specific tool.
Becker et al. (2006)	Eligibility	Instrument not main focus	Instruments used as predictors/covariates rather than the focus of validation.
Calle-Ramirez et al. (2020)	Eligibility	Language	Full-text version not available in English.
Audrain-McGovern et al. (2006)	Eligibility	Instrument not main focus	Self-control measure used for prediction; not a behavioural audit focus.
Marcus (1999)	Eligibility	No psychometric evaluation	Lacks sufficient report of psychometric properties (e.g., reliability/validity).
Ruiselov (2000)	Eligibility	No psychometric evaluation	Abstract only; insufficient data to assess instrument properties.
Adams (1995)	Eligibility	No psychometric evaluation	Copyright notice/snippet only; no psychometric data provided.
Paschall (2001)	Eligibility	No psychometric evaluation	Lacks empirical psychometric data for extracted instruments.
Kohn (1979)	Eligibility	No psychometric evaluation	Abstract only; cannot verify quantitative validation design.
Senna (1974)	Eligibility	No psychometric evaluation	Descriptive study focused on academic investment; no instrument validation.
Shapiro (1997)	Eligibility	No psychometric evaluation	Lacks evidence of psychometric evaluation or instrument-centric focus.
Miller (1980)	Eligibility	No psychometric evaluation	Abstract only; insufficient detail on psychometric properties.
Connor (Unknown)	Eligibility	No psychometric evaluation	Lacks novel psychometric data; focus on prevalence/predictors.
DeLisi (2008)	Eligibility	No psychometric evaluation	Focus on institutional misconduct relationships, not scale validation.
Quay (1977)	Eligibility	No psychometric evaluation	Descriptive overview lacking quantitative psychometric metrics.
Banárová (2024)	Eligibility	Instrument not main focus	Focus on cross-variable relationships; validation not primary objective.
Dickson (2005)	Eligibility	No psychometric evaluation	Focus on prevalence/risk factors without reporting instrument properties.
Adams (1997)	Eligibility	No psychometric evaluation	Copyright snippet only; no psychometric evidence available.
Ribner (1979)	Eligibility	No psychometric evaluation	Abstract only; insufficient psychometric detail.
Bilenberg (1999)	Eligibility	No psychometric evaluation	Snippet only; lacks report of psychometric coefficients.
Andreea (2010)	Eligibility	No psychometric evaluation	Journal overview/TOC; no specific instrument validation data.
Tardivo (2022)	Eligibility	Language	Full-text not available in English (Portuguese primary).

Author (Year)	Exclusion Stage	Primary Rationale for Rejection	Specific Tier/Reason
McGuire (2000)	Eligibility	Language	Non-English text; uninterpretable for extraction.
Maya (2023)	Eligibility	Language	Full-text available in Spanish only.
Albores-Gallo (2007)	Eligibility	Language	Predominantly Spanish; insufficient English data for extraction.
Hoffmann (2020)	Eligibility	Not peer-reviewed	Unclear peer-review status at the time of assessment.
Crowley (1993)	Eligibility	Not peer-reviewed	Cannot verify peer-review status in available source.
Kemp (2024)	Eligibility	Unclear peer-review	Manuscript version; peer-review status not finalized.
Beld (2017)	Eligibility	Wrong Construct	Focus on classroom climate (Tier 5/9); not antisocial behavior.
Mehari (2025)	Eligibility	Wrong Construct	Focus on "Hope" and positive development; not antisocial (Tier 10).
Santos (2018)	Eligibility	Wrong Construct	Measures Early Maladaptive Schemas; not behavioral audit (Tier 8).
Vilte (2016)	Eligibility	Wrong Construct	Focus on Empathy (Tier 6); no primary antisocial behavior focus.
Huitink (2011)	Eligibility	Wrong Construct	Staff behavior measures; not adolescent antisocial behavior.
Albiero (2009)	Eligibility	Wrong Construct	Focus on Empathy (Tier 6) rather than conduct outcomes.
Bachman (1986)	Eligibility	Wrong Construct	Focus on Type A behavior (Tier 8); lacks behavioral conduct focus.
Coulter (2024)	Eligibility	Wrong Construct	Focus on Positive Youth Development; not antisocial (Tier 10).
Almiro (2023)	Eligibility	Wrong Construct	Apperception test focused on broad functioning; lacks behavioral audit items.
Benesch (2014)	Eligibility	Wrong Population	Child sample (Age 6-12); outside adolescent range.
Kazdin (1987)	Eligibility	Wrong Population	Child sample (Age 6-12); does not meet primary age range.
Archer (2010)	Eligibility	Wrong Population	Adult male sample; focus on hypermasculinity.
Weiss (2002)	Eligibility	Wrong Population	Primary school sample (Age 8-14); not primarily adolescent.
Broxholme (2003)	Eligibility	Wrong Population	Adult focus with intellectual disabilities.
Maxwell (2007)	Eligibility	Wrong Population	Sample not specific to adolescents; broad general population.
Levy (2013)	Eligibility	Wrong Population	Clinic-referred children; age not specific to adolescence.
Archer (1997)	Eligibility	Wrong Population	Undergraduate student samples (Adult).
Bai (2017)	Eligibility	Wrong Population	Early childhood sample (Age 6-10).
Alder (2007)	Eligibility	Wrong Population	Adult sample (Age 18-61).
Ruiz-Pamies (2014)	Eligibility	Wrong Population	Broad age range (Adult focus); pilot used undergraduates.
Casat (1999)	Eligibility	Wrong Population	Elementary school children (Mage = 8.3).

Author (Year)	Exclusion Stage	Primary Rationale for Rejection	Specific Tier/Reason
Margari (2005)	Eligibility	Wrong Population	Adult psychiatric inpatients (Age 18-65).
van Baardewijk (2008)	Eligibility	Wrong Population	Children (Age 9-12).
Fite (2024)	Eligibility	Wrong Population	Children (Age 9-12).
Stein (1995)	Eligibility	Wrong Population	Study on parents of children with ADHD.
Nederlof (2009)	Eligibility	Wrong Population	Undergraduate sample (Mage = 19.8).
Rhee (2015)	Eligibility	Wrong Population	Children (Age 2-12).
Tsakanikos (2011)	Eligibility	Wrong Population	Adults with intellectual disabilities (Age 18-74).
Delforterie (2020)	Eligibility	Wrong Population	Adult clients with intellectual disabilities.
Lengua (2001)	Eligibility	Wrong Population	Primarily focuses on child symptomatology; age range not specific to adolescents.
Hawes (2013)	Eligibility	Wrong Population	Children (Age 6-12).
Burbidge (2010)	Eligibility	Wrong Population	General population with intellectual disability; not adolescent-specific.
Webster (2014)	Eligibility	Wrong Population	Adult college student focus.
Denson (2006)	Eligibility	Wrong Population	General population sample; not primarily adolescent.
Hukkelberg (2016)	Eligibility	Wrong Population	Children (Age 2-12).
Pekarik (1976)	Eligibility	Wrong Population	Child sample; specific adolescent focus not established.
Van Leeuwen (2006)	Eligibility	Wrong Population	Young children (Age 4-8).
Gomez (Unknown)	Eligibility	Missing Year	Bibliographic record incomplete; year of publication unavailable.
Pechorro et al. (2016)	Taxonomic	A2. Psychometric/General Risk	Analyzes MMPI-A (broad personality instrument); focus on factor structure refinement rather than core behavioral antisociality.
Pelham et al. (1992)	Taxonomic	L2. Specific Symptom Measures	RS-DBD based on DSM ODD/ADHD symptom clusters; fails to capture the broad antisocial construct.
Nieuwenhuijzen (2005)	Taxonomic	B2. Core Underlying Constructs	BITI measures irrational thoughts (cognitive trait); psychological mechanism, not a behavioral measure.
Carlini et al. (2014)	Taxonomic	I1. Substance Use/Risk Behaviour	CTCYS focuses on risk factors for substance use; falls under mixed problem/general risk surveys.
Harpur et al. (2014)	Taxonomic	C2. Anger/Hostility/Irritability	Examines trait anger as a predictor; aggression is the outcome, not the instrument's core construct.
Kuperminc et al. (2001)	Taxonomic	B2. Core Underlying Constructs	ABAS development focused on antisocial beliefs/thinking (mechanisms), not behavioral acts.

Author (Year)	Exclusion Stage	Primary Rationale for Rejection	Specific Tier/Reason
Kuperminc et al. (2004)	Taxonomic	B2. Core Underlying Constructs	Psychometric development of ABAS (cognitions); underlying mechanism underlying behavior.
Tangney et al. (1996)	Taxonomic	C2. Anger/Hostility/Irritability	ARI focuses on anger arousal and management responses; primarily an emotional state measure.
Stams et al. (2009)	Taxonomic	J1. Empathy/Moral Cognition	SRM-SFO measures moral judgment/reasoning; defined as a psychological mechanism (Tier 6).
Stringaris et al. (2012)	Taxonomic	L2. Specific Symptom Measures	Focuses on breaking down ODD into irritability/defiance clusters; not broad antisocial behavior.
Steinhausen et al. (2007)	Taxonomic	L2. Specific Symptom Measures	DAYAS primarily assesses specific symptom clusters of ADHD and ODD.
D'Alessio et al. (2014)	Taxonomic	K1. Family/Parenting Measures	FPEB measures parenting and peer-relation quality; primarily contextual influences.
Mowbray et al. (2014)	Taxonomic	K1. Family/Parenting Measures	TIRE measures within-family environment and parental criticism; family functioning construct.
Martel et al. (2010)	Taxonomic	L2. Specific Symptom Measures	Focuses on modeling ADHD/ODD symptom dimensions rather than behavioral antisociality.
Vittengle et al. (2011)	Taxonomic	K2. Peer/Social Context	CYCIS measures classroom-specific incivility; context-bound peer environment measure.
Contreras et al. (2012)	Taxonomic	K1. Family/Parenting Measures	TXP designed to evaluate parenting practices and control; family process construct.
Rivera-Rivera (2007)	Taxonomic	E2. Dating/Sexual Aggression	M-CTS adapted specifically for dating violence; specialized aggression domain.
Sánchez-Hernández (2020)	Taxonomic	E2. Dating/Sexual Aggression	CyDAV-T developed for cyber dating violence; intimate partner aggression focus.
Grietens et al. (2004)	Taxonomic	I2. Clinical/Treatment Context	CBRF-A designed for daily staff ratings of change within an inpatient treatment process.
Moral de la Rubia (2011)	Taxonomic	C2. Anger/Hostility/Irritability	ML-STAXI-MS core construct is the emotion of anger and its expression.
Mazurek et al. (2019)	Taxonomic	L1. Developmental Condition Scales	OS-CBS developed for "challenging behavior" specific to ASD/DD populations.
Casas et al. (2007)	Taxonomic	K1. Family/Parenting Measures	SCIH analyzes sibling interactions; context-bound relational aggression within the family.
Hervas et al. (2022)	Taxonomic	L1. Developmental Condition Scales	ACB scale co-developed for the autistic community to assess ASD-specific concerning behaviors.
Zhang et al. (2011)	Taxonomic	I1. Substance Use/Risk Behaviour	HBICA is a health-risk inventory including suicide, substance use, and unprotected sex.
Brotman et al. (2017)	Taxonomic	C2. Anger/Hostility/Irritability	TAPIS specifically designed to measure tonic/phasic irritability (angry mood).
Wolfe et al. (2001)	Taxonomic	E2. Dating/Sexual Aggression	TeDAM expansion of CADRI; focuses on traditional indicators of dating violence.

Author (Year)	Exclusion Stage	Primary Rationale for Rejection	Specific Tier/Reason
Henly et al. (1988)	Taxonomic	I1. Substance Use/Risk Behaviour	PESQ explicitly designed as an alcohol and drug abuse screening scale.
De Clercq et al. (2006)	Taxonomic	A2. Psychometric/General Risk	DAPP-BQ-A is a broad personality pathology scale; lacks specific behavioral audit focus.
Gázquez et al. (2015)	Taxonomic	K2. Peer/Social Context	CONVIVE evaluates behaviors performed/received in the school climate context.
Espelage et al. (2003)	Taxonomic	E2. Dating/Sexual Aggression	J-SOAP-II is a specialized forensic tool for assessing sexual reoffense risk.
Hussong et al. (2005)	Taxonomic	I1. Substance Use/Risk Behaviour	Evaluates CTC-F7 familial factors specifically within alcohol-intoxicated samples.
Dierckx et al. (2012)	Taxonomic	K1. Family/Parenting Measures	Measures mothers' relational schemas and family affective attitudes; parenting construct.
Srinath et al. (2013)	Taxonomic	C2. Anger/Hostility/Irritability	Development of an anger scale based on behavioral/verbal/mental levels of emotion.
Bullis et al. (1994)	Taxonomic	L1. Developmental Condition Scales	BULLIS developed and validated specifically for youth with emotional/behavioral disorders (EBD).
Waldron et al. (1996)	Taxonomic	I1. Substance Use/Risk Behaviour	PADDI structured interview for co-occurring substance use and mental health disorders.
Gomez et al. (2016)	Taxonomic	L2. Specific Symptom Measures	CADBI focused on specific DSM-IV symptom clusters of ODD and ADHD.
Calvete et al. (2012)	Taxonomic	J1. Empathy/Moral Cognition	Measures justification of violence and grandiosity schemas (moral cognition mechanisms).
Govender et al. (2013)	Taxonomic	I1. Substance Use/Risk Behaviour	Uses CTC sub-scales to predict tobacco, alcohol, and marijuana use (substance focus).
Gibbs et al. (1995)	Taxonomic	J1. Empathy/Moral Cognition	HIT Questionnaire measures antisocial cognitive distortions (moral/cognitive mechanisms).
Lansford et al. (2012)	Taxonomic	I2. Clinical/Treatment Context	TOPS-A measures responses to problematic situations in secure institutional/correctional care.
Arthur et al. (2002)	Taxonomic	I1. Substance Use/Risk Behaviour	CTCYS measures risk/protective factors for substance abuse across ecological domains.
Perez et al. (2016)	Taxonomic	J1. Empathy/Moral Cognition	MMDS measures Bandura’s mechanisms of moral disengagement (cognitive restructuring).
Werner et al. (2003)	Taxonomic	K2. Peer/Social Context	Participant Role Behavior Scales measure peer group dynamics (assisting, outsider roles).
Van der Helm (2011)	Taxonomic	K2. Peer/Social Context	PIRY measures peer interactions and social climate within residential youth care.
Ghandour et al. (2012)	Taxonomic	B2. Core Underlying Constructs	ABIS measures impulsivity (psychological trait) rather than behavioral antisocial acts.
Dickens et al. (2011)	Taxonomic	I2. Clinical/Treatment Context	Designed to predict treatment dropout/recidivism in specialized residential sex-offender programs.

Author (Year)	Exclusion Stage	Primary Rationale for Rejection	Specific Tier/Reason
Costin et al. (2011)	Taxonomic	L1. Developmental Condition Scales	BPI-01 validated specifically for children/adolescents with intellectual disabilities.
Uji et al. (2016)	Taxonomic	C2. Anger/Hostility/Irritability	Cross-cultural adaptation of Affective Reactivity Index (tonic irritability measure).
Ilic et al. (2011)	Taxonomic	A2. Psychometric/General Risk	ACE-ASF measures adverse childhood experiences/trauma exposure (environmental risk).
Zhang et al. (2012)	Taxonomic	L1. Developmental Condition Scales	BPI-01 translated/validated for aggression assessment in ID/Specialized populations.
Handel et al. (2011)	Taxonomic	A2. Psychometric/General Risk	Evaluation of MMPI-A (broad psychopathology tool) subscales in a forensic population.
Tayhan et al. (2021)	Taxonomic	B2. Core Underlying Constructs	BIS-Brief measures the primary construct of impulsivity; mechanism, not behavioral act.
Warshofsky (2012)	Taxonomic	L2. Specific Symptom Measures	VADTRS primarily designed to measure ADHD symptom clusters (inattention/hyperactivity).
Stanton et al. (1996)	Taxonomic	C2. Anger/Hostility/Irritability	Adolescent version of Cook-Medley scale measures cynical hostility/mistrust as a trait.
Sánchez-Hernández (2016)	Taxonomic	I1. Substance Use/Risk Behaviour	Systematic review of CTCYS; focused on risk/protective factors for mixed problem behaviors.
Winters et al. (1993)	Taxonomic	I1. Substance Use/Risk Behaviour	DAB validated within a sample explicitly hospitalized for substance abuse.
Sullivan et al. (2004)	Taxonomic	A2. Psychometric/General Risk	ETV scale measures traumatic event exposure/witnessing rather than antisocial behavior.
Olofsdotter (2019)	Taxonomic	L2. Specific Symptom Measures	Investigate teacher-rated symptoms for specific DSM clusters of ODD and CD.
Douglas et al. (2014)	Taxonomic	I2. Clinical/Treatment Context	RSTI-SR evaluates psycho-legal maturity and treatment amenability; specialized context.
Fite et al. (2012)	Taxonomic	B2. Core Underlying Constructs	Study examines self-control as a moderator; psychological mechanism (executive function).
Gilead et al. (2014)	Taxonomic	E2. Dating/Sexual Aggression	Secondary analysis focused on sexual aggression and victimization in school contexts.
Pechorro et al. (2016b)	Taxonomic	A2. Psychometric/General Risk	SAS-A-SF measures social anxiety (internalizing); unrelated to antisocial phenotypes.
Hoza et al. (2005)	Taxonomic	J2. General Social/Prosocial	CBCL Peer Rejection subscale measures social difficulties/immaturity (inverse construct).
Conners et al. (2011)	Taxonomic	L2. Specific Symptom Measures	Conners 3(R) focus is ADHD and associated specific behavioral symptom clusters.
Muncer et al. (1997)	Taxonomic	B2. Core Underlying Constructs	EXPAGG measures social representation/attributions (cognitive schemas) of aggression.

Author (Year)	Exclusion Stage	Primary Rationale for Rejection	Specific Tier/Reason
Caputo et al. (2011)	Taxonomic	E2. Dating/Sexual Aggression	Validates J-SOAP-II/ERASOR risk tools for specialized sexual offense recidivism.
Heras et al. (2020)	Taxonomic	E2. Dating/Sexual Aggression	ADV scale assesses attitudes justifying aggression in dating relationships.
Adams et al. (1999)	Taxonomic	L2. Specific Symptom Measures	ABC explicitly designed as a measure of ADHD symptomatology.
Rojahn et al. (2012)	Taxonomic	L1. Developmental Condition Scales	BPI-01 developed for aggressive/destructive behavior in developmental disabilities.
Garcia-Sanchez (2017)	Taxonomic	E2. Dating/Sexual Aggression	CDVI measures online dating abuse/partner violence (specialized domain).
Prior et al. (2010)	Taxonomic	L1. Developmental Condition Scales	DBC designed to assess behavioral disturbance specifically in individuals with MR/ID.
Andersson (2011)	Taxonomic	I1. Substance Use/Risk Behaviour	EuroADAD/ADAD designed for drug use and associated psychosocial problems.
Mewborn (1981)	Taxonomic	L1. Developmental Condition Scales	HK/MBD questionnaire focused on developmental learning/hyperactivity symptoms.
Nas et al. (2016)	Taxonomic	J1. Empathy/Moral Cognition	Validates HIT Questionnaire in Bangladeshi culture (cognitive distortion mechanism).
Shields et al. (2001)	Taxonomic	J2. General Social/Prosocial	Multiple-choice measure of avoidance and assertion (adaptive/prosocial responses).
Törnhage (2016)	Taxonomic	I1. Substance Use/Risk Behaviour	ADAD assesses drug abuse severity and treatment need; substance-centric focus.
Wang et al. (2012)	Taxonomic	I1. Substance Use/Risk Behaviour	AIUQ measures internet use and problematic screen-based risky behavior.
Visontay et al. (2021)	Taxonomic	J1. Empathy/Moral Cognition	MDS measures moral disengagement proneness; mechanism, not behavior.
Schmidt et al. (2011)	Taxonomic	J2. General Social/Prosocial	SAI-Y measures personal strengths and self-concept; adaptive functioning construct.
Öhrvik et al. (2014)	Taxonomic	K1. Family/Parenting Measures	APQ measures parenting practices (discipline, monitoring); family process construct.
Oosterlaan et al. (2010)	Taxonomic	L2. Specific Symptom Measures	Focuses on validity of ADHD and HKD symptom domains (inattention/hyperactivity).
Snyder et al. (2012)	Taxonomic	I2. Clinical/Treatment Context	TPTO:YAB assesses therapist judgments of treatment success in MST programs.
Kivitz et al. (2012)	Taxonomic	J1. Empathy/Moral Cognition	BSGQ measures emotions of shame and guilt (moral cognition/affective mechanisms).
Reijntjes et al. (2016)	Taxonomic	J2. General Social/Prosocial	Measures resistance to peer pressure (adaptive skill/protective factor).
Van Voorhis (1994)	Taxonomic	I2. Clinical/Treatment Context	HRSQ-YO measures situational antecedents to offense rather than the behavior.
Abad et al. (2012)	Taxonomic	A2. Psychometric/General Risk	JS NEO-A60 measures Big Five personality domains; pure trait inventory.
Gázquez et al. (2016)	Taxonomic	K2. Peer/Social Context	Deviant Peers Scale assesses behavior in peers; environmental/contextual measure.

Author (Year)	Exclusion Stage	Primary Rationale for Rejection	Specific Tier/Reason
Coccaro et al. (2011)	Taxonomic	L2. Specific Symptom Measures	M-IED developed for diagnosis of Intermittent Explosive Disorder (symptom cluster).
Arbisi et al. (1998)	Taxonomic	A2. Psychometric/General Risk	Analyzes content scales of MMPI-A (general psychopathology tool).
Jiménez et al. (2011)	Taxonomic	K1. Family/Parenting Measures	SLEQ measures past victimization/trauma; environmental/historical risk factor.

Note. Taxonomic categories (A2–L2) correspond to the exclusion rationales defined in the Taxonomic Framework (Table S3). Exclusions primarily distinguish instruments measuring behavioral outcomes (included) from those measuring underlying mechanisms, specialized domains, or environmental contexts (excluded). "Wrong Population" denotes samples primarily outside the 12–18y developmental window. "No psychometric evaluation" indicates studies utilizing a tool descriptively without providing novel quantitative validation (e.g., factor structure, internal consistency).

Table S5.

Studies Excluded During the Consolidation Phase: Population Age Refinement (n=69)

Author (Year)	Exclusion Stage	Primary Rationale for Rejection	Specific Tier / Reason
Achenbach (1987)	Consolidation	Wrong Age Group	Boys aged 6–11 included; mean age below adolescent window.
Beck et al. (2009)	Consolidation	Wrong Age Group	Sample range 5.0–18.0 years (Mean 12.9); included pre-latency children.
Becker et al. (2012)	Consolidation	Wrong Age Group	Nationally representative sample includes children aged 3–11.
Van Baardewijk (2010)	Consolidation	Wrong Age Group	Includes children as young as 8; mean age below target threshold.
Essau et al. (2011)	Consolidation	Wrong Age Group	Mean age 11.5 (*SD* = 3.3*)*; substantial portion of sample is pre-adolescent.
Pechorro et al. (2015)	Consolidation	Wrong Age Group	Sample includes young adults up to 22 years (*M* = 17.7, *SD* = 1.7*)*.
Bögels et al. (2003)	Consolidation	Wrong Age Group	Sample includes children aged 1.5–5 and 6–11.
Lynam et al. (2005)	Consolidation	Wrong Age Group	Mean age 13.2 (*SD* = 2.4*)*; range starts at 9 years (Tier 1 violation).
Epstein et al. (1998)	Consolidation	Wrong Age Group	Mean age 11.2; focuses on early puberty/latency (Age 8–16).
Achenbach (1981)	Consolidation	Wrong Age Group	Includes 4-year-olds; fails to isolate 12–18y adolescent data.
Shaffer et al. (2004)	Consolidation	Wrong Age Group	T1 data focuses on children typically aged 6–11.
Goodman (2001)	Consolidation	Wrong Age Group	Broad 3–16 year old sample; cannot isolate adolescent psychometrics.
Meltzer et al. (2003)	Consolidation	Wrong Age Group	5–16 year old range; sample mean heavily weighted toward childhood.
Korn et al. (2001)	Consolidation	Wrong Age Group	Grade-based sample (Grades 1–10); includes elementary children.
Ebesutani et al. (2010)	Consolidation	Wrong Age Group	Mean age 11.4 (*SD* = 2.5*)*; range starts at 6.5 years.
Blank et al. (2011)	Consolidation	Wrong Age Group	Broad 4–18 year old sample (Mean age 11).
Ebesutani et al. (2011)	Consolidation	Wrong Age Group	Includes children as young as 7; focuses on younger youths.
Finning et al. (2017)	Consolidation	Wrong Age Group	Focuses exclusively on the 11-year-old wave of the MCS.
Sawyer et al. (2001)	Consolidation	Wrong Age Group	6–17 year old range; lacks distinct adolescent-only psychometrics.
Collett et al. (2003)	Consolidation	Wrong Age Group	Focuses on 6–12 year old children.
Pechorro et al. (2014)	Consolidation	Wrong Age Group	Portuguese translation tested on ages 7–17 (Mean 13.3).
Hussey et al. (2012)	Consolidation	Wrong Age Group	Sample includes 8-year-old survivors; mean age below 12.
Pineda et al. (2015)	Consolidation	Wrong Age Group	Colombian sample includes 10-year-olds; range 10–23y.
Kamphaus et al. (2015)	Consolidation	Wrong Age Group	Focus on Preschool and Child Clinical scales.
Sullivan et al. (2000)	Consolidation	Wrong Age Group	Mean age 14.4; but range starts at 9 years (Tier 1 violation).
Merrell et al. (2002)	Consolidation	Wrong Age Group	Mean age 11.2; broad 5–18y validation sample.
Eyjolfsdottir (2016)	Consolidation	Wrong Age Group	Icelandic sample includes 5-year-olds; range 5–16y.
Borg (2004)	Consolidation	Wrong Age Group	Maltese version includes children as young as 3 years.

Author (Year)	Exclusion Stage	Primary Rationale for Rejection	Specific Tier / Reason
Little et al. (2003b)	Consolidation	Wrong Age Group	Focuses on elementary school children.
Schmidt et al. (2008)	Consolidation	Wrong Age Group	Sample mean age 20.4 (*SD* = 1.6*)*; emerging adult sample.
Almiro et al. (2015)	Consolidation	Wrong Age Group	Sample includes young adults up to age 22.
Panzani et al. (2020)	Consolidation	Wrong Age Group	Includes 6-year-old children with ID.
Mongut et al. (2019)	Consolidation	Wrong Age Group	Broad 5–18 year old sample (CANS validation).
Krueger et al. (2011)	Consolidation	Wrong Age Group	Focus on late adolescence to young adulthood (MTI longitudinal).
Frick et al. (2015)	Consolidation	Wrong Age Group	Clinician-rating focus for ages 3–15 (*M* = 7.4*)*.
Pechorro et al. (2011)	Consolidation	Wrong Age Group	Sample range 13–21 years (*M* = 16.8*)*.
Calvete et al. (2015)	Consolidation	Wrong Age Group	Sample includes adults up to 24 years (*M* = 16.2*)*.
Price et al. (2001)	Consolidation	Wrong Age Group	Rasch analysis includes children from age 4.
Lyons et al. (2002)	Consolidation	Wrong Age Group	Broad 5–18 year old mental health sample.
Little et al. (2003c)	Consolidation	Wrong Age Group	Emerging adulthood focus; sample range not primarily adolescent.
Girdziene et al. (2003)	Consolidation	Wrong Age Group	Focuses on the 7–11 year old Lithuanian subgroup.
Zubrick et al. (2014)	Consolidation	Wrong Age Group	4–17 year old range; includes urban aboriginal children.
Pechorro et al. (2016)	Consolidation	Wrong Age Group	Parent-reported version for ages 6–15 (*M* = 11*)*.
Price et al. (2002)	Consolidation	Wrong Age Group	Brief clinical measure for ages 6–18.
Price et al. (1999)	Consolidation	Wrong Age Group	NCBRF typical version; mean age 10.0.
Muris et al. (2003)	Consolidation	Wrong Age Group	Community sample of 9–15 year olds.
Kersten et al. (2016)	Consolidation	Wrong Age Group	Includes the 3-year-old wave of the Millennium Cohort Study.
Frick et al. (2016)	Consolidation	Wrong Age Group	Parent-report PCS for ages 6–17.
Merrell et al. (2012)	Consolidation	Wrong Age Group	IRT sample includes 5-year-old children.
Price et al. (2008)	Consolidation	Wrong Age Group	BRACHA development includes children from age 3.5.
Olofsdotter et al. (2016)	Consolidation	Wrong Age Group	6–18 year old Swedish clinical sample.
Offord et al. (1992)	Consolidation	Wrong Age Group	Focus on 6–11 year old children in OCHS.
Becker et al. (2015)	Consolidation	Wrong Age Group	Children of prisoners sample (Mean age 11.2).
Lynam et al. (2007)	Consolidation	Wrong Age Group	Meta-analysis/Review including age 11.
Kersten et al. (2012)	Consolidation	Wrong Age Group	Includes age 3 wave in longitudinal cohort.
Pineda et al. (2016)	Consolidation	Wrong Age Group	9–15 year olds (Early adolescent focus).
Price et al. (2011)	Consolidation	Wrong Age Group	Retrospective R-MOAS for ages 5–17 (*M* = 11.2*)*.
Pechorro et al. (2017)	Consolidation	Wrong Age Group	Detained sample includes adults up to 24 years.
Price et al. (2016)	Consolidation	Wrong Age Group	Parent-reported ICU for ages 6–15 (*M* = 11*)*.
Nas et al. (2012)	Consolidation	Wrong Age Group	Sample range 18–35 years (*M* = 20.7*)*.
Raine et al. (2006)	Consolidation	Wrong Age Group	Mean age 16.1; but range includes emerging adults.

Author (Year)	Exclusion Stage	Primary Rationale for Rejection	Specific Tier / Reason
Fan et al. (2014)	Consolidation	Wrong Age Group	Chinese translation tested on ages 3–17.
Van der Loo et al. (2015)	Consolidation	Wrong Age Group	Refugee sample includes adults up to 26 years (*M* = 15.7*)*.
Pineda et al. (2013)	Consolidation	Wrong Age Group	9–16 year old Chilean sample (*M* = 11.5*)*.
Price et al. (2014)	Consolidation	Wrong Age Group	PCS self-report for ages 9–15 (*M* = 12.2*)*.
Ribeiro et al. (2009)	Eligibility	Duplicate	Secondary citation of previously extracted dataset.
Krohn et al. (2010)	Eligibility	Duplicate	Duplicate record of variety/frequency analysis.
Smith et al. (2013)	Eligibility	Duplicate	Secondary report of same validation sample.
Gomez et al. (2016)	Eligibility	Duplicate	Redundant record of agreement analysis.

Note. Rationales for exclusion are mapped Inclusion Exclusion criteria in Table S2. Duplicates include secondary reports of previously extracted primary datasets. Wrong Population refers to samples outside the target adolescent window (12–18 years). An a priori buffer of ±2 years was applied; studies were retained if participants fell within the 10–20 year range. Studies excluded under Population Age at the Consolidation stage were those confirmed via full text as primarily focused on childhood (<10 years) or emerging adulthood (>20 years).

Table S6.

Data Extraction Framework: Parameter Definitions and Psychometric Variables

Variable Name	Description
SECTION 1 — STUDY METADATA	
Study_ID	Unique identifier assigned to each included study.
First_Author_Year	Citation-style reference (e.g., Smith, 2021).
Country	Country in which the sample was collected.
Sample_Size	Total number of participants included in analyses.
Age_Range	Age span of the sample (e.g., 12–17 years).
Population_Type	Community, clinical, justice-involved, school, etc.
Setting	Context of data collection (e.g., school, clinic, detention).
SECTION 2 — INSTRUMENT METADATA	
Instrument_Name	Full name of the instrument evaluated.
Instrument_Acronym	Short form (e.g., ICU, YPI, RPQ).
Subscale_Name	Specific subscale examined (if applicable).
Construct_Domain	Target construct (e.g., aggression, CU traits, conduct).
Number_of_Items	Total number of items in the scale or subscale.
Informant	Type of rater (self, parent, teacher, clinician).
Language	Language in which the instrument was administered.
Is_Adapted_Version	Indicates if translated or culturally adapted (Yes/No).
SECTION 3 — RELIABILITY	
Alpha	Cronbach's alpha coefficient.
Omega	McDonald's omega coefficient.
TestRetest_ICC	Test–retest intraclass correlation coefficient.
InterRater_ICC	Inter-rater reliability coefficient (if applicable).
SECTION 4 — STRUCTURAL VALIDITY	
Structural_Analysis_Type	Type of structural analysis conducted (CFA, EFA, Rasch, IRT).
Intended_Factor_Structure	Theoretically expected factor model.
Extracted_Factor_Structure	Empirically derived factor model.
RMSEA	Root Mean Square Error of Approximation.
CFI	Comparative Fit Index.
TLI	Tucker–Lewis Index.
Measurement_Invariance	Whether invariance was tested and supported (Yes/No).
SECTION 5 — VALIDITY EVIDENCE	
Convergent_Validity_Effect	Effect size with theoretically related constructs.
Divergent_Validity_Effect	Effect size with unrelated constructs.
Criterion_Validity_Effect	Association with gold-standard or external criteria.
Predictive_Validity_Effect	Predictive effect (AUC, OR, β) for future antisocial outcomes.
Known_Groups_Validity	Whether known-group differences were tested (Yes/No).
Hypotheses_Confirmed	Proportion of a priori hypotheses supported.
Validity_Notes	Additional validity-related notes (free text).
SECTION 6 — FEASIBILITY	
Administration_Time	Estimated time required to complete the measure.
Training_Required	Whether specialised training is required to administer.

Variable Name	Description
Cost_or_License	Free, paid, or restricted licensing requirements.
SECTION 7 — COSMIN RISK OF BIAS RATINGS	
COSMIN_RoB_StructuralValidity	Quality rating for structural validity.
COSMIN_RoB_InternalConsistency	Quality rating for internal consistency evidence.
COSMIN_RoB_Reliability	Quality rating for reliability (test–retest/inter-rater).
COSMIN_RoB_Validity	Quality rating for validity evidence.
COSMIN_Global_Rating	Overall COSMIN evidence rating for the instrument.
SECTION 8 — OVERALL INSTRUMENT ASSESSMENT	
Overall_Psychometric_Quality	High, moderate, or low overall quality.
Strengths	Summary of psychometric strengths.
Limitations	Summary of limitations.
Suitability_for_Adolescents	Appropriateness for adolescent populations (Yes/No/Mixed).

Note. This table summarises the 45 core variables used for standardised data extraction based on COSMIN (COnsensus-based Standards for the selection of health Measurement INstruments) methodology. Variables are grouped into eight categories capturing essential information for evaluating reliability, structural validity, various forms of validity evidence, feasibility, and risk of bias across instruments measuring antisocial behaviour in adolescents. This reduced extraction framework improves efficiency, facilitates comparison across instruments, and aligns with best practices in psychometric systematic reviews.

Table S7.

Evaluative Criteria for Rating the Quality of Measurement Property Outcomes

Measurement property	Outcome rating	Criteria applied
Internal consistency	Sufficient (+)	Cronbach's α or McDonald's ω ≥ .70 *and* evidence of unidimensionality or acceptable structural validity
	Insufficient (–)	α or ω < .70
	Indeterminate (?)	No reliability estimate reported or no evidence of unidimensionality
Structural validity (CFA)	Sufficient (+)	CFI/TLI ≥ .90 and RMSEA ≤ .08 (SRMR ≤ .08 where reported)
	Insufficient (–)	CFI/TLI < .90 or RMSEA > .10
	Indeterminate (?)	Insufficient reporting of model fit indices
Structural validity (EFA)	Sufficient (+)	Clear factor structure supported by appropriate methods (e.g. parallel analysis), theoretical coherence, and primary loadings ≥ .40
	Insufficient (–)	Unstable or poorly interpretable factor solution
	Indeterminate (?)	Insufficient detail to evaluate factor structure
Test–retest reliability	Sufficient (+)	ICC ≥ .70
	Insufficient (–)	ICC < .70
	Indeterminate (?)	No stability estimate reported
Construct validity (hypothesis testing)	Sufficient (+)	≥ 75% of pre-specified hypotheses confirmed
	Insufficient (–)	< 75% of hypotheses confirmed
	Indeterminate (?)	Hypotheses not specified
Criterion validity	Sufficient (+)	Correlation ≥ .70 with appropriate criterion or acceptable classification indices
	Insufficient (–)	Correlation < .70
	Indeterminate (?)	No clear criterion or insufficient information
Measurement invariance	Sufficient (+)	At least metric invariance supported
	Insufficient (–)	Invariance rejected
	Indeterminate (?)	Invariance not tested or insufficiently reported

Note. Outcome ratings were assigned in accordance with COSMIN guidelines for the evaluation of measurement properties. Ratings reflect the adequacy of the reported evidence for each psychometric property rather than the overall quality of individual studies. Structural validity was evaluated holistically based on the totality of reported fit indices and factor-analytic evidence, rather than strict adherence to single cut-off values. Internal consistency ratings were interpreted only when supported by evidence of unidimensionality. When insufficient information was reported to permit judgement, outcomes were rated as indeterminate (?).

Table S8.

Full Psychometric Extraction and Methodological Quality for 125 Included Instrument Variants

Study	Variant	Inf	Items	Sample (*N*)	α	Method	Fit Indices	Inv/Validity	R: SV	R: IC	R: Rel	R: V	Cat.
Warren et al. (1989)	AABC	Staff	21	Clinical / 55	0.91	NR		Con V 0.54 / Div V Strong	Inadequate	Very Good	Very Good	Very Good	B
Howard (2009)	AAS	Youth	NR	Forensic / 101	0.96	CFA	CFI 0.93 / TLI 0.92 / RMSEA 0.07	Crit V Strong	Very Good	Very Good	NR	Very Good	D
Burt & Donnellan (2009)	ABS	Youth	16	National / 12714	0.74	IRT		MI DTF	Adequate	Adequate	Inadequate	Very Good	B
Pritchard et al. (2011)	AFS	Parent	33	Community / 279	0.9	PCA		Con V 0.78 / Div V 0.09	Inadequate	Very Good	Adequate	Very Good	A
Pechorro et al. (2021)	AHSRD	Youth	10	Risk / 200	0.81	CFA	CFI 0.99 / TLI 0.99 / RMSEA 0.07	Con V 0.65 / Div V Strong	Very Good	Adequate	Very Good	Very Good	B
Richter et al. (2011)	APBFS	Youth	15	Community / 1602	0.91	EFA / CFA	CFI 0.95 / TLI 0.95 / RMSEA 0.069	MI Weak / Div V 0.8	Very Good	Very Good	Inadequate	Adequate	C
Frick & Hare (2001)	APSD	Multi (Y/P/M)	17–20	Diverse; N=5,909	0.52–0.84	CFA / IRT / ESEM	CFI: .84–.94; RMSEA: .04–.06	MI: Supported; Conv: Mod/Strong; Pred: ASPD/Crime	Inadequate	Very Good	Inadequate	Very Good	B
Loney & Frick (2003)	APSD-SR	Youth	17–20	Diverse; N=7,255	0.42–0.77	CFA / EFA / Regress	CFI: .87–.95; RMSEA: .04–.08	Conv: Mod/Strong; Crit: Yes; Discrim: Foren vs Comm	Inadequate	Inadequate	Inadequate	Adequate	B
Buss & Perry (1992)	AQ	Youth / Multi	12–29	Diverse; N=4,674	0.62–0.92	CFA / ESEM	CFI: .92–.95; RMSEA: .06–.08	MI: Partial (Gender/Culture); Conv: Mod/Strong; Crit: Yes	Inadequate	Inadequate	Inadequate	Adequate	C
Morales-Vives et al. (2014)	AQ-PA	Youth	28	Community / 1458	0.85	CFA		MI Yes / Con V 0.65 / Crit V Yes	Very Good	Very Good	Inadequate	Very Good	C
Bryant & Smith (2001)	AQ-SF	Youth	12	Diverse; N=1,640	0.81–0.92	CFA	CFI: .90–.95; RMSEA: .06–.08	MI: Yes (Gender/Culture); Conv: Mod (0.55–0.60); Crit: Yes	Adequate	Very Good	Inadequate	Very Good	C
Orpinas & Frankowski (2001)	AS	Youth	11	Community / 1223	0.89	PCA		Con V 0.65 / Crit V Yes / Pred V Predict	Inadequate	Very Good	Adequate	Very Good	C
Mak (1993)	ASRDS	Youth	24–34	Community; N=1,936	0.90–0.91	CFA / EFA	CFI: .93–.96;	Conv: Mod/High; Div: Yes; Crit: Predictive	Adequate	Very Good	Very Good	Very Good	B

Study	Variant	Inf	Items	Sample (*N*)	α	Method	Fit Indices	Inv/Validity	R: SV	R: IC	R: Rel	R: V	Cat.
Webster et al. (2014)	BAQ	Youth	12	Community / 3996	0.81	EFA / CFA / IRT	RMSEA: .04–.07 CFI 0.986 / TLI 0.963 / RMSEA 0.05	Con V 0.7 / Div V 0.3 / Crit V Predict / Pred V Exp Predict	Very Good	Very Good	Very Good	NR	C
Reynolds & Kamphaus (2004)	BASC PRS-A	Parent	48	Community / 200	0.86	CFA	CFI 0.94 / TLI 0.93 / RMSEA 0.05	MI Metric / Con V Yes / Crit V Yes	Very Good	Very Good	NR	Very Good	A
Kamphaus & Reynolds (2015)	BASC-3 BESS	Youth	30	Community / 400	0.89	ROC		Con V 0.6 / Crit V Yes	Inadequate	Very Good	NR	Very Good	A
Reynolds & Kamphaus (2004)	BASC-SRP	Youth	186	Residential / 970	0.91	CFA	CFI 0.91 / TLI 0.9 / RMSEA 0.1	Con V 0.7 / Div V 0.2	Adequate	Very Good	Adequate	Doubtful	A
Context Dependent (Study Specific)	Battery	Youth	NR	Forensic / 154	0.9	CFA		MI Yes / Con V 0.3 / Crit V Yes	Adequate	Very Good	Inadequate	Very Good	H
Thomas et al. (2017)	BCS-A	Youth	64	Community / 3044	0.97	CFA	CFI 0.9 / TLI 0.89 / RMSEA 0.05	MI Yes / Con V 0.6 / Crit V Yes	Very Good	Very Good	Inadequate	Very Good	E
Buss & Durkee (1957)	BDHI	Youth	75	Community / 106	0.9	Correl / T-test		Con V 0.52 / Crit V Yes	Inadequate	Very Good	Very Good	Very Good	C
Bryant & Smith (2001)	BPAQ-SF	Youth	12	Diverse; N=3,396	0.89	CFA / Rasch	CFI: .98; TLI: .97; RMSEA: .06	MI: Yes; Conv: Mod; Pred: Crime Severity	Adequate	Very Good	Very Good	Very Good	C
Quay (1977)	BPC	Staff	55	Forensic / 50	0.86	NR			Inadequate	Very Good	Adequate	NR	B
Bergen (2001)	BQAB	Youth	33	Community / 2752	0.91	PCA		Con V 0.45	Inadequate	Very Good	Adequate	Very Good	B
Halliday et al. (2016)	BullyHARM	Youth	21	Community / 397	0.93	CFA	CFI 0.992 / TLI 0.99 / RMSEA 0.023		Very Good	Very Good	Inadequate	NR	E
Abdul Bari & Wan Mohd (2020)	Bullying Scale	Youth	18	Community / 677	0.88	EFA / PCA			Inadequate	Very Good	Inadequate	Adequate	E
Buss & Warren (2000)	BWAQ	Youth	29	Community / 1018	0.93	EFA / CFA		Con V 0.97 / Crit V SAVRS	Adequate	Very Good	NR	Very Good	C
Recent (e.g., 2023 Study)	BWAQ-ML	Youth	15	Community / 1018	0.89	CFA / ML	CFI 0.95 / TLI 0.94 / RMSEA 0.04	Con V 0.97 / Crit V SAVRS	Very Good	Very Good	NR	Very Good	C
Conners (1997)	CASS:S	Youth	27	Clinical / 240	0.73	PCA	CFI 0.87	Con V 0.52	Inadequate	Adequate	Inadequate	Adequate	A

Study	Variant	Inf	Items	Sample (*N*)	α	Method	Fit Indices	Inv/Validity	R: SV	R: IC	R: Rel	R: V	Cat.
Achenbach (1991)	CBCL (Parent)	Parent	113-120	Diverse; N=1,923	0.65–0.94	CFA / ANOVA	CFI: .92–.97; RMSEA: .04–.12	MI: Supported (AA Sample); Conv: Mod; Crit: Strong	Inadequate	Very Good	Inadequate	Very Good	A
Yang (2000)	CBCL-C	Parent	118	Community / 854	0.96	CFA	CFI 0.97 / TLI 0.97 / RMSEA 0.04	MI Supported / Con V 0.51 / Crit V Yes	Very Good	Very Good	Very Good	Very Good	A
Achenbach (1991)	CBCL-T	Teacher	113	Clinical / 145	0.88	CFA	CFI 0.96 / RMSEA 0.09	MI Failed / Con V 0.78 / Pred V 0.39	Adequate	Adequate	Inadequate	Adequate	A
Calvete et al. (2010)	CBQ	Parent	23	Multi / 218	0.86	CFA	CFI 0.987 / TLI 0.986 / RMSEA 0.07	Con V 0.5 / Div V 0.15	Adequate	Very Good	Inadequate	Very Good	E
Antoniadou et al. (2016)	CBVEQ-G	Youth	24	Community / 1097	0.89	CFA	CFI 0.97 / TLI 0.97 / RMSEA 0.031	MI Yes / Con V Strong	Very Good	Very Good	Inadequate	Very Good	E
Gomez (2007)	CD Trait	Multi	NR	Referred / 203	0.82	CFA/MTMM		Con V Yes / Crit V Yes	Adequate	Very Good	NR	Very Good	B
Pineda et al. (2003)	CD-CL	Multi	15	Forensic / 138	0.86	NR		Con V High / Crit V Discrim	Inadequate	Very Good	Adequate	Very Good	B
Romera et al. (2018)	CGQ-A	Youth	16	Community / 1431	0.96	CFA	CFI 1 / TLI 1 / RMSEA 0.016	Con V Strong	Very Good	Very Good	Inadequate	Very Good	E
Tuvblad et al. (2014)	Composite	Multi	NR	Twin / 660	0.85	CFA	CFI 0.9 / TLI 0.89 / RMSEA 0.05	MI Yes / Con V Yes / Crit V Yes	Very Good	Very Good	Adequate	Very Good	G
Contreras et al. (2012)	CPAQ	Youth	20–24	Community; N=3,074	0.76–0.91	CFA	CFI: .97–.99; RMSEA: .04–.05	MI: Yes; Conv: 0.77; 1- & 4-Factor Models (Mother/Father)	Very Good	Very Good	Inadequate	Very Good	F
Recent (e.g., Covid-19 Update)	CPAQ-R	Youth	25	Community / 1244	0.93	ESEM / CFA	CFI 0.994 / TLI 0.99 / RMSEA 0.019	Con V 0.58	Very Good	NR	NR	NR	F
Lynam (1997)	CPS	Youth	66	Risk / 279	0.9	Regress		Con V Moderate / Crit V Yes / Pred V Agg Predict	Inadequate	Very Good	NR	Adequate	C
Lynam et al. (2005)	CPS-SR	Youth	38	Community / 818	0.9	CFA/EFA	CFI 0.91 / TLI 0.9 / RMSEA 0.06	MI Yes / Con V 0.8 / Div V -0.2 / Crit V 0.5	Very Good	Very Good	Adequate	Very Good	G

Study	Variant	Inf	Items	Sample (*N*)	α	Method	Fit Indices	Inv/Validity	R: SV	R: IC	R: Rel	R: V	Cat.
Gázquez et al. (2016)	CPV-Q	Multi (Y/P)	27–28	Diverse; N=2,538	0.55–0.93	CFA / EFA	CFI: .97–1.0; RMSEA: .00–.04	4-Factor Structure; MI: Partial; Conv: Mod	Adequate	Doubtful	Inadequate	Adequate	F
Pechorro et al. (2015)	CQA	Youth	21	Community / 508	0.8	EFA			Adequate	Adequate	Inadequate	Adequate	B
Straus (1979)	CTS	Youth	19	Community / 106	0.84	Correl / T-test		Con V 0.52 / Crit V Yes	Inadequate	Very Good	NR	Very Good	C
Graf et al. (2012)	Cyber-Quest	Youth	25	Community / 2148	0.82	CFA / SEM	CFI 0.975 / TLI 0.962 / RMSEA 0.039	Con V High / Pred V Predict	Adequate	Adequate	NR	NR	E
Wachs et al. (2020)	Cyberhate	Youth	20	Community / 1005	0.63	CFA	CFI 0.93 / TLI 0.91 / RMSEA 0.06	Con V Low	Adequate	Adequate	Inadequate	Adequate	E
Shook et al. (2014)	DARAS	Youth	NR	Community / 275	0.86	CFA			Adequate	Very Good	NR	Adequate	C
Henry et al. (2012)	DAS	Youth	33	Forensic / 190	0.96	EFA		Con V Strong / Div V Strong / Crit V Strong	Adequate	Very Good	Inadequate	Very Good	B
Wichstrøm et al. (1991)	Delinq	Youth	NR	Community / 10000	NR	EFA		Crit V Variety Index	Adequate	Inadequate	Inadequate	Inadequate	B
López-Romero et al. (2019)	DISBA	Youth	29	Community / 1446	0.94	EFA		Con V 0.76 / Div V 0.3	Adequate	Very Good	Adequate	Very Good	B
LeBuffe et al. (1994)	DSMD	Parent/Youth	110	Clinical / 108	0.8	Correl		Con V 0.8 / Div V 0.16 / Crit V K-SADS / Pred V 76% Acc	Inadequate	Very Good	Very Good	Very Good	A
Ortega-Ruiz et al. (2016)	EBIP-Q	Youth	5	Community / 1777	0.74	CFA	CFI 0.96 / TLI 0.95 / RMSEA 0.054	MI Yes / Con V Strong / Crit V Yes	Very Good	Adequate	Inadequate	Very Good	E
Zeman et al. (2001)	EBS	Youth	NR	Community / 258	Satisfactory	EFA			Adequate	Adequate	NR	Adequate	C
Eyberg (1980)	ECBI	Parent	35	Clinical / 167	0.96	EFA		Con V 0.74 / Div V 0.6 / Crit V Discrim	Adequate	Very Good	Inadequate	Very Good	B
Del Rey et al. (2015)	ECIP-Q	Youth	5–22	Community; N=4,006	0.74–0.93	CFA / EFA	CFI: .96–.99; RMSEA: .04–.05	MI: Yes; Conv: Strong; 2-Factor (Agg/Victim)	Very Good	Adequate	Inadequate	Adequate	E
Klein et al. (2001)	Eurogang	Youth	4	Community / 5083	NR	Chi-Square		Con V 0.5 / Crit V No	Inadequate	Inadequate	Inadequate	Inadequate	B
Shaw et al. (2013)	FBS	Youth	36	Community / 2512	0.92	CFA	CFI 0.95 / TLI 0.95 / RMSEA 0.048	MI Yes / Con V 0.6 / Crit V Yes	Very Good	Very Good	Inadequate	Very Good	E

Study	Variant	Inf	Items	Sample (*N*)	α	Method	Fit Indices	Inv/Validity	R: SV	R: IC	R: Rel	R: V	Cat.
Shaw et al. (2013)	FBS-P	Youth	18	Community / 2512	0.92	CFA	CFI 0.95 / TLI 0.95 / RMSEA 0.048	MI Yes / Con V 0.6 / Crit V Yes	Very Good	Very Good	Inadequate	Very Good	E
Shaw et al. (2013)	FBS-V	Youth	18	Community / 2512	0.92	CFA	CFI 0.95 / TLI 0.95 / RMSEA 0.048	MI Yes / Con V 0.53 / Crit V Yes	Very Good	Very Good	Inadequate	Very Good	E
Basto-Pereira et al. (2015)	Frequency	Youth	17	Community / 1292	0.77	Regress		Crit V No	Inadequate	Adequate	Doubtful	Doubtful	B
Recent Ad-hoc	HAT	Youth	12	Adolescents / 402	0.91	EFA / CFA		Con V 0.44 / Pred V Predict	Adequate	Very Good	Inadequate	Adequate	E
Frick (2004)	ICU	Multi (Y/P)	10–24	Diverse; N=15,410	0.76–0.87	CFA / IRT / EFA	CFI: .81–.96; RMSEA: .00–.10	MI: Supported (Gender/Culture); Conv: Mod/Strong; Pred: Agg/Crime	Inadequate	Inadequate	Inadequate	Inadequate	G
Pechorro et al. (2017)	ICU-10	Youth	10	Forensic / 221	0.79	CFA	CFI 0.96 / TLI 0.95 / RMSEA 0.05	Con V 0.5 / Crit V Yes	Very Good	Adequate	Inadequate	Very Good	G
Wang et al. (2019)	ICU-11	Youth	11	Diverse; N=1,414	0.76–0.81	CFA	CFI: .96–.97; RMSEA: .03–.04	MI: Supported; Conv: Mod (0.60); 2-Factor (Callous/Uncaring)	Very Good	Adequate	Inadequate	Very Good	G
Pechorro et al. (2017)	ICU-12	Youth	12	Diverse; N=1,380	0.76–0.81	CFA	CFI: .95–.97; RMSEA: .05–.07	MI: Supported; Conv: Mod (0.45) to High (0.94); 3-Factor Model	Very Good	Adequate	Inadequate	Very Good	G
Frick (2004)	ICU-SR	Youth	24	Clinical / 67	0.72	EFA		Con V 0.38 / Div V -0.36 / Crit V No / Pred V No	Adequate	Adequate	Adequate	Inadequate	G
Spanish Adaptation	INCA	Youth	33	Forensic / 288	0.92	CFA	CFI 0.92 / TLI 0.9 / RMSEA 0.05	MI Yes / Con V Strong / Crit V Yes	Very Good	Very Good	Inadequate	Very Good	G
Mexican Adaptation	INCA-M	Youth	33	Community / 586	0.92	CFA	CFI 0.9 / TLI 0.89 / RMSEA 0.048	Con V 0.56 / Crit V Yes	Very Good	Very Good	Inadequate	Very Good	G
Barratt et al. (1999)	IPAS	Youth	20	Clinical / 120	0.9	PCA		Con V Strong / Pred V Subtyping	Very Good	Very Good	NR	Very Good	D
Polman et al. (2009)	IRPA	Youth	NR	Community / 957	Good	CFA			Adequate	Adequate	NR	Adequate	D
Enzmann et al. (2010)	ISRD	Youth	NR	Community / 5083	0.82	Correl		Con V 0.6 / Crit V Yes	Inadequate	Very Good	Inadequate	Very Good	B
Jesness (2003)	Jesness	Self-report	68	Risk / Clinical / 597	0.91	CFA / IRT		Con V 0.65 / Crit V Judicialization	Very Good	Very Good	Adequate	Very Good	B

Study	Variant	Inf	Items	Sample (*N*)	α	Method	Fit Indices	Inv/Validity	R: SV	R: IC	R: Rel	R: V	Cat.
Little et al. (2003)	LAI	Youth	NR	Risk / 289	0.9	CFA	CFI 0.9 / TLI 0.88 / RMSEA 0.05		Adequate	Very Good	NR	Very Good	D
Brown et al. (1979)	LHA	Youth	9	Community / NR	Sound	NR			Inadequate	Adequate	NR	Adequate	C
Millon (1993)	MACI	Clinician	20	Clinical / 227	0.93	CFA	CFI 0.96 / TLI 0.95 / RMSEA 0.043	Con V 0.78 / Div V Yes / Crit V Yes	Very Good	Very Good	Inadequate	Very Good	G
Grisso et al. (2001)	MAYSI-2	Youth	52	Forensic; N=4,385	0.74–0.91	PCA / Rasch	NR	MI: Yes (DIF); Conv: Strong; Crit: Yes	Inadequate	Adequate	Inadequate	Adequate	H
Matson et al. (1983)	MESSY	Youth	62	Community / 634	0.91	EFA		Con V 0.72 / Div V 0.61	Adequate	Very Good	Inadequate	Very Good	A
Paquin et al. (2014)	MIA	Youth	113	Community / 1443	0.85	CFA	CFI 0.86 / RMSEA 0.03	Con V Strong / Div V Adequate	Very Good	Very Good	Inadequate	Very Good	A
Archer et al. (2016)	MMPI	Youth	58	Forensic / 1053	0.75	CCA / Correl		MI Yes / Con V Yes / Div V Yes / Crit V Yes / Pred V Primary/Sec	Inadequate	Adequate	Inadequate	Adequate	G
Seeley et al. (2002)	OADP-CDS	Self-report	6	Community / 1709	NR	ROC Analysis		Crit V ASPD Predict	Inadequate	Adequate	Adequate	Very Good	B
Vreugdenhil et al. (2006)	OC	Staff	30	Forensic / 154	0.87	CFA		MI Yes / Con V 0.3 / Crit V Yes	Adequate	Very Good	Adequate	Very Good	H
Achenbach-mapped	ODD/CD	Youth	NR	Clinical / 658	0.75	CFA / ESEM	CFI 0.956 / TLI 0.944 / RMSEA 0.058	MI Scalar / Con V 0.46 / Div V 0.27	Very Good	Adequate	Inadequate	Very Good	B
Sukhodolsky et al. (2016)	OMS	Parent	24	Community/Clinical / 431	0.9	CFA	CFI 0.96 / TLI 0.95 / RMSEA 0.05	MI Yes / Con V Strong / Div V Strong / Crit V Discrim	Very Good	Very Good	Adequate	Very Good	C
Barnoski (2004)	PACT	Clinician	52	Forensic / 1540	NR	ROC / Surv		Crit V 0.73 / Pred V 0.73	Inadequate	NR	NR	Very Good	H
Pechorro et al. (2016)	PAS	Youth	16	Forensic / 259	0.81	EFA		Con V 0.56 / Crit V 0.71	Adequate	Very Good	Inadequate	Very Good	G
Farrell et al. (2000)	PBFS	Youth	35	Community / 5532	NR	CFA	CFI 0.971 / TLI 0.968 / RMSEA 0.021	MI Strong / Con V Strong / Div V Strong / Crit V Strong	Very Good	Inadequate	Very Good	Very Good	B
Forth et al. (2003)	PCL:YV	Clinician	18–20	Forensic; N=1,288	0.79–0.89	CFA / IRT / ROC	CFI: .90–.95; RMSEA: .06–.07	Conv: Mod (0.37–0.47); Pred: High (AUC .76)	Inadequate	Very Good	Very Good	Very Good	G

Study	Variant	Inf	Items	Sample (*N*)	α	Method	Fit Indices	Inv/Validity	R: SV	R: IC	R: Rel	R: V	Cat.
Marsee et al. (2011)	PCS-20	Youth	20	Forensic / 192	0.88	CFA	CFI 0.91 / RMSEA 0.06	MI Yes	Adequate	Adequate	NR	Adequate	D
Wright & Li (2013)	Peer-EA	Peer	2	Adolescents / 573	NR	NR		Con V 0.29 / Pred V Predict	NA	NA	Adequate	Adequate	E
Salekin (2016)	PSCD	Youth	12–29	Diverse; N=4,503	0.87–0.91	CFA	CFI: .91–.96; RMSEA: .04–.06	MI: Supported; Conv: Strong (.80); Hierarchical/Bifactor structure	Very Good	Very Good	Inadequate	Very Good	G
Salekin (2016)	PSCD-P	Parent	12–29	Community; N=2,711	0.90–0.94	CFA	CFI: .95–.97; TLI: .94–.96; RMSEA: .05	MI: Supported; Conv: Strong; 4-Factor Model	Very Good	Very Good	Inadequate	Very Good	G
Salekin (2016)	PSCD-SV	Youth	29	Community / 522	0.87	CFA	CFI 0.9 / TLI 0.89 / RMSEA 0.05	MI Yes / Con V 0.7 / Div V 0.15 / Crit V 0.35	Very Good	Very Good	Adequate	Very Good	G
Cala et al. (2022)	RAQ	Youth	32	Community / 355	0.92	CFA	CFI 0.968 / TLI 0.963 / RMSEA 0.041	MI Yes / Con V 0.77 / Crit V Yes	Very Good	Very Good	Inadequate	Very Good	F
Quay (1983)	RBPC	Parent	89	Clinical / 144	NR	EFA			Adequate	Inadequate	Inadequate	NR	B
De Los Reyes & Prinstein (2004)	RPEQ	Youth	28	Community / 1320	0.88	CFA	CFI 0.973 / RMSEA 0.035	MI Partial / Con V 0.6	Very Good	Very Good	Doubtful	Adequate	E
Raine et al. (2006)	RPQ	Youth	22–23	Diverse; N=16,695	0.84–0.92	CFA / EFA / PCA / LCA	CFI: .85–.99; RMSEA: .03–.07	MI: Supported (Gender/Culture); Conv: Mod (0.50); 2-Factor Structure	Inadequate	Adequate	Inadequate	Adequate	D
Goodman-linked	S-SBQ	Teacher	22	Community / 239	0.9	ROC		Con V 0.44 / Div V Low / Crit V DISC / Pred V OR > 5.0	Inadequate	Very Good	Adequate	Very Good	A
Thornberg & Jungert (2013)	SBPS	Youth	11	Community / 1600	0.93	CFA	CFI 0.881 / RMSEA 0.035	MI Scalar / Con V 0.64	Very Good	Very Good	Doubtful	Adequate	E
Thornberg & Jungert (2013)	SBVS	Youth	11	Community / 1600	0.92	CFA	CFI 0.983 / RMSEA 0.03	MI Scalar / Con V 0.72	Very Good	Very Good	Doubtful	Adequate	E
Giannakopoulos et al. (2009)	SCPS	Youth	88	Community / 853	0.87	PCA		Con V 0.7	Inadequate	Very Good	Very Good	Adequate	A
Koskelainen et al. (2000)	SCS	Youth	80	Multi / 1356	0.92	ANOVA		Crit V Criterion / Pred V Large ES	Inadequate	Very Good	Inadequate	Very Good	A
Muris et al. (2004)	SDQ-DP	Youth	15	Community / 1732	0.87	CFA	CFI 0.94 / TLI 0.93 /	Con V 0.75 / Pred V Predict	Very Good	Very Good	Inadequate	Very Good	A

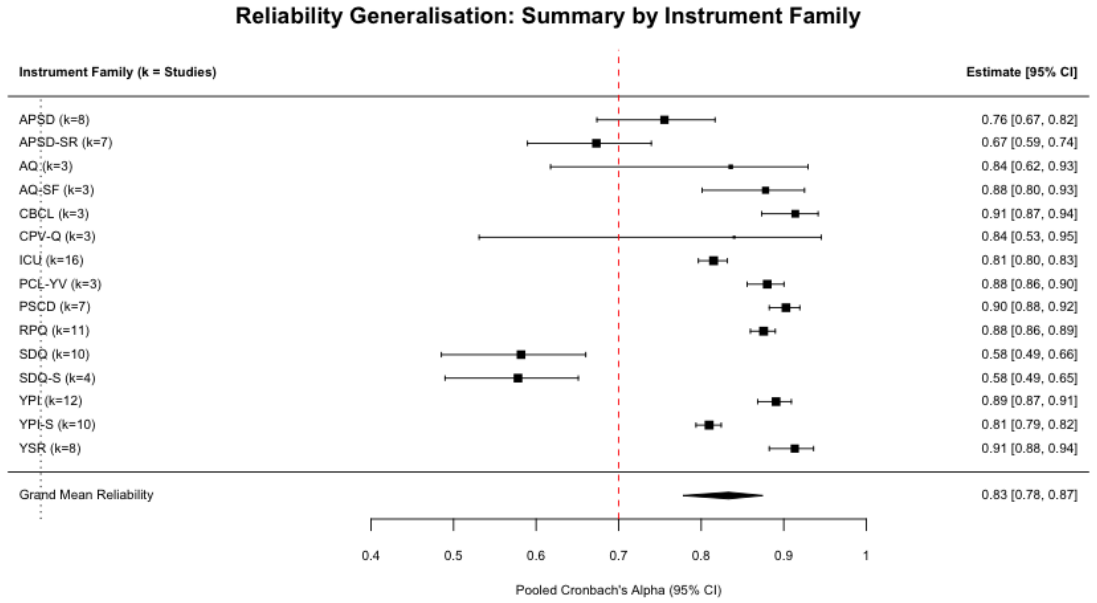
Study	Variant	Inf	Items	Sample (*N*)	α	Method	Fit Indices	Inv/Validity	R: SV	R: IC	R: Rel	R: V	Cat.
Goodman (1999)	SDQ-IS	Youth	10	Students / 400	0.77	ROC	RMSEA 0.04	Con V 0.55 / Crit V Yes	Inadequate	Adequate	Inadequate	Very Good	A
Goodman (1997)	SDQ-S	Youth	20–25	Diverse; N=91,154	0.17–0.77	CFA / EFA / Rasch / PCA	CFI: .82–.96; RMSEA: .02–.06	MI: Partial to Strict; Conv: Mod/Strong; 3, 4, 5 & 6-factor models	Inadequate	Inadequate	Inadequate	Inadequate	A
Various (e.g., 22-item)	Short-CBCL	Multi	57	Forensic / 1749	0.76	Regress		Con V 0.4 / Crit V Yes	Inadequate	Adequate	NR	Adequate	A
Elliott et al. (1985)	SRD	Youth	NR	Risk / 412	0.74	CFA / ESEM	CFI 0.98 / TLI 0.96 / RMSEA 0.07	Con V 0.7 / Div V 0.19	Very Good	Adequate	Adequate	Very Good	B
Walker & Severson (1992)	SSBD	Teacher	10	Clinical / 123	0.97	Correl		Con V 0.78 / Div V 0.81 / Crit V Discrim / Pred V ODR scores	Inadequate	Very Good	Very Good	Very Good	A
Burt & Donnellan (2009)	STAB	Youth	32	Community / 1293	0.79	EFA / CFA	CFI 0.941 / RMSEA 0.053	MI Strict / Con V 0.6 / Div V 0.4 / Crit V Predict / Pred V Predict	Very Good	Adequate	Adequate	Very Good	B
Spielberger (1973)	STAIC	Youth	10	Forensic / 309	0.91	NR		Con V Yes / Div V Yes / Crit V Yes	NA	Very Good	Inadequate	Very Good	G
Achenbach (1991)	TRF	Teacher	113	Refugee / 486	0.91	CFA		MI Implicit / Con V 0.81 / Div V 0.2 / Pred V 26% Cut	Very Good	Very Good	Very Good	Very Good	A
Yang (2000)	TRF-C	Teacher	113	Community / 162	0.97	CFA	CFI 0.97 / TLI 0.97 / RMSEA 0.04	MI Supported / Con V 0.51 / Crit V Yes	Very Good	Very Good	Very Good	Very Good	A
Patrick (2010)	TriPM	Youth	58	Community / 1540	0.83	CFA / Reg	CFI 0.9 / TLI 0.88 / RMSEA 0.06	MI Yes / Con V 0.6 / Div V Yes / Crit V Predict	Very Good	Adequate	Adequate	Very Good	G
Norwegian National Data	Ungdata	Youth	6	Community / 297102	0.79	CFA / IRT	CFI 0.985 / TLI 0.982 / RMSEA 0.038	MI Scalar	Very Good	Adequate	Inadequate	Very Good	B
Basto-Pereira et al. (2015)	Variety	Youth	17	Community / 1292	0.87	Regress		Crit V Yes / Pred V Yes	Inadequate	Very Good	Adequate	Adequate	B
Hagler (2016)	WBI	Youth	20	Community / 2444	0.89	CFA	CFI 0.96 / TLI 0.95 / RMSEA 0.07	Con V Strong / Div V Strong / Pred V 4-yr change	Very Good	Very Good	Adequate	Very Good	A
Gadow & Sprafkin (1999)	YEBS	Youth	12	Community / 894	0.81	EFA / CFA	CFI 0.98 / TLI 0.02 /	Crit V 0.78	Very Good	Very Good	Inadequate	Very Good	A

Study	Variant	Inf	Items	Sample (*N*)	α	Method	Fit Indices	Inv/Validity	R: SV	R: IC	R: Rel	R: V	Cat.
Hagler (2016)	YEPS	Youth	10	Community / 396	0.89	EFA	RMSEA 0.06 CFI 0.97 / TLI 0.95 / RMSEA 0.06	Con V 0.74 / Div V 0.33	Very Good	Very Good	Inadequate	Inadequate	A
Gadow & Sprafkin (1999)	YI-4	Youth	180	Clinical / 239	0.77	NR		Con V Strong / Div V Moderate / Crit V AUC .84	Inadequate	Adequate	Adequate	Adequate	A
Hagler (2016)	YIEPS	Youth	44	Students / 400	0.95	ROC		Con V 0.6 / Crit V Yes	Inadequate	Very Good	Inadequate	Very Good	A
Hoge & Andrews (2002)	YLS/CMI	Clinician	42	Forensic; N=4,598	0.77	ROC / CFA / Surv	CFI: .90; TLI: .89; RMSEA: .05	MI: Supported (Sex); Pred: Mod/High (0.66–0.76)	Inadequate	Adequate	Very Good	Very Good	H
Andershed et al. (2002)	YPI	Youth	50	Diverse; N=5,876	0.72–0.92	CFA / EFA / ROC	CFI: .89–.96; RMSEA: .04–.06	MI: Supported; Conv: Mod/High; 3-Factor Structure	Inadequate	Adequate	Inadequate	Adequate	G
van Baardewijk et al. (2010)	YPI-S	Youth	18	Diverse; N=7,248	0.79–0.88	CFA / Reg / ROC	CFI: .89–.95; RMSEA: .04–.06	MI: Supported; Conv: Mod/Strong; 3-Factor (Inter/Aff/Beh)	Inadequate	Adequate	Inadequate	Adequate	G
Achenbach (1991)	YSR	Youth	33–103	Diverse; N=10,473	0.80–0.96	CFA / IRT / PCA	CFI: .93–.97; TLI: .92–.97; RMSEA: .03–.07	MI: Partial; Conv: High (0.94); Broadband Stable; Narrowband cross-loads	Inadequate	Inadequate	Inadequate	Adequate	A
Achenbach (1991)	YSR-AB	Youth	18	Forensic / 310	0.91	CFA	CFI 0.93 / TLI 0.92 / RMSEA 0.07	MI Yes / Con V Strong	Very Good	Very Good	NR	Very Good	C
Murray et al. (2019)	ZBBS	Youth	10	Community / 1304	NR	CFA	CFI 0.956 / RMSEA 0.044	MI Partial / Con V 0.32 / Div V Prosocial / Crit V Predict	Adequate	Inadequate	Inadequate	Adequate	E

Note. This table continues for 125 rows in total. Inf = Informant; N = sample size; α = Cronbach's alpha; Method = structural validation method; Inv = invariance; MI = measurement invariance; Con V = convergent validity; Div V = divergent validity; Crit V = criterion validity; Pred V = predictive validity. R: = COSMIN Risk of Bias ratings (SV = Structural Validity; IC = Internal Consistency; Rel = Reliability; V = Validity). Cat. = Taxonomic classification according to the 12-tier taxonomic audit (see Table S3 for definitions). NR = Not Reported.

Figure S1.

Forest Plot of Pooled Internal Consistency Coefficients for Key Adolescent Antisocial Behaviour Instrument Families



Note. This forest plot displays the pooled Cronbach's alpha (α) estimates derived from the random-effects reliability generalisation meta-analysis for instrument families with k≥3 independent samples. The size of each diamond is proportional to the instrument's weight in the meta-analysis, reflecting its precision based on total sample size and variance. Horizontal bars indicate 95% Confidence Intervals (CIs). The vertical dashed line represents the α=.70"reliability floor." Instruments are ordered to illustrate the "Psychometric Premium," where strongly aligned multidimensional tools (e.g., BASC, CBCL, PSCD) demonstrate superior and more stable precision (pooled α =.88) compared to the lower reliability and higher variability observed in partially aligned scales (e.g., SDQ Conduct Problems, ICU).

Chapter 3. Bridging Chapter: From Systematic Appraisal to Empirical Evaluation

The systematic review presented in Chapter 2 provided a structured evaluation of instruments designed to assess adolescent antisocial behaviour. Applying a sequential 12-tier taxonomic audit followed by alignment with the four DSM-5 Conduct Disorder (CD) domains substantially reduced the measurement landscape. Although many instruments claim to assess antisocial behaviour under diverse labels and with varying degrees of psychometric support, only a small subset demonstrated both adequate behavioural coverage and sufficient methodological robustness under COSMIN and GRADE criteria. Variability was particularly evident in structural validation practices, with frequent reliance on principal component analysis and limited reporting of measurement invariance. Furthermore, instruments demonstrating comprehensive domain coverage often required extended administration time, whereas briefer measures prioritised feasibility at the potential expense of construct breadth.

Taken together, the review clarified an important pattern within the field: although numerous tools exist and many report internal consistency estimates, relatively few simultaneously demonstrate comprehensive coverage of the full developmental breadth of antisocial behaviour — ranging from lower-frequency or milder acts such as truancy and minor rule-breaking to more serious behaviours including property destruction, weapon carrying, and acts with potential legal consequences — alongside rigorous latent structural evaluation within a developmentally appropriate format. This tension between behavioural scope, methodological rigour, and practical feasibility emerged as a central challenge in adolescent antisocial behaviour assessment.

The findings of the review therefore carry important methodological implications. If adolescent antisocial behaviour is conceptualised as a dimensional construct spanning a gradient of behavioural severity, then instruments intended for screening or research use must

demonstrate more than internal consistency alone. Robust evaluation requires evidence of structural validity using latent modelling approaches, appropriate reliability estimation, and examination of whether measurement properties operate equivalently across key demographic and developmental groups. In adolescence, where behavioural expression may shift across age and context, limited testing of dimensional structure and measurement invariance constrains confidence in comparisons across gender, age bands, or longitudinal waves. Under such conditions, observed differences may reflect variation in measurement properties rather than underlying behavioural change.

In addition to the instruments identified in the systematic review, several large-scale longitudinal cohort studies have incorporated concise behavioural indices intended to capture antisocial acts aligned with DSM-5 Conduct Disorder domains. These measures were typically developed for inclusion within broader epidemiological investigations and therefore prioritised brevity and administrative feasibility. However, pragmatic inclusion within cohort designs does not necessarily imply that such instruments have undergone comprehensive psychometric evaluation using contemporary latent modelling approaches. In many cases, detailed examination of dimensional structure, multi-group measurement invariance, and diagnostic classification performance has not been systematically reported.

The Antisocial Behaviour Questionnaire (ABQ), implemented within the ROOTS and NeuroScience in Psychiatry Network (NSPN) cohorts, provides an opportunity to examine these issues empirically. As a concise behavioural index mapped to the four DSM-5 Conduct Disorder domains and administered across a broad developmental span, it offers a test case for whether a brief instrument can satisfy the standards of structural validity, developmental stability, and diagnostic coherence identified in the preceding review. A comprehensive psychometric evaluation of its dimensional structure, measurement invariance, reliability, and external validity is therefore warranted.

This thesis portfolio proceeds in two complementary stages. Chapter 2 presented a systematic evaluation of adolescent antisocial behaviour instruments, integrating a 12-tier taxonomic audit with DSM-5 Conduct Disorder domain alignment and COSMIN-based psychometric appraisal. Building on those clarified standards, Chapter 4 provides a focused psychometric evaluation of the ABQ using large UK community cohorts spanning adolescence and emerging adulthood. Together, the two papers aim to advance the precision and coherence of antisocial behaviour measurement by integrating systematic field appraisal with targeted empirical validation.

Chapter 4. Psychometric Evaluation, Diagnostic Utility, and Measurement Invariance of the Antisocial Behaviours Questionnaire (ABQ) in two UK Adolescent Cohorts

Punita Panchal[1], Kenny Chiu[1], Sharon A S Neufeld[2]

[1]Department of Clinical Psychology and Psychological Therapies, Norwich Medical School, University of East Anglia, Norwich Research Park, Norwich NR4 7TJ, UK

[2]Department of Psychiatry, University of Cambridge, Cambridge CB2 8AH, UK

This manuscript has been formatted for submission to Psychological Assessment (PA). Abstract word count: 230. Manuscript word count based on PA guidelines (not including title page, abstract, public significance statement, and supplementary material): 9796

Acknowledgements

SN was supported by the Wellcome Trust Early Career Award (226392/Z/22/Z). All research in the Department of Psychiatry at Cambridge University benefits from the National Institute for Health Research (NIHR) Cambridge Biomedical Research Centre.

Disclaimer/Declaration

The views and opinions expressed therein are those of the authors and do not necessarily reflect those of the NIHR, NHS or the Department of Health and Social Care.

Author Note

We have no conflicts of interest to disclose.

Correspondence concerning this article should be addressed to Sharon A S Neufeld, [2]Department of Psychiatry, University of Cambridge, Cambridge CB2 8AH, UK Email: sasn2@medschl.cam.ac.uk

Abstract

Antisocial behaviours (AB) are critical markers of developmental risk, yet their assessment is often hindered by a lack of concise tools that align with clinical criteria. Many brief screeners fail to meet the reliability thresholds required for clinical interpretation. The Antisocial Behaviours Questionnaire (ABQ) addresses this void by providing an 8-item index mapped directly to the four DSM-5 Conduct Disorder domains: aggression, destruction, deceit, and rule violations. We evaluated the ABQ's psychometric properties in two large UK community cohorts (ROOTS, $N = 1{,}153$; NSPN, $N = 2{,}378$) coverage ages 14–24. Using binary-appropriate weighted least squares estimation, exploratory and confirmatory factor analyses supported a robust unidimensional structure. The single-factor solution explained 56% of common variance, with near-perfect fit indices (CFI ≥ .999; RMSEA ≤ .013) and strong factor loadings (.48–.94). Reliability was exceptionally strong by $\omega = .93$ ($\alpha = .73$). Measurement invariance supported comparisons across gender and age, including the transition from adolescence to young adulthood. Construct validity was confirmed through convergence with psychopathic traits ($r = .41$–$.53$) and prospective prediction of substance misuse. Diagnostic utility for Conduct Disorder was excellent, with a threshold of ≥ 2 behaviours optimizing the balance between sensitivity and specificity. As a low-burden, open-access screening tool, the ABQ is a developmentally sensitive tool for high-precision assessment of antisocial behaviours in community samples.

Keywords: Antisocial behaviour; Adolescence; Confirmatory Factor Analysis; Measurement Invariance. **Public Significance Statement:** This study validates the Antisocial Behaviours Questionnaire (ABQ) as a brief, reliable tool for identifying antisocial behaviours in young people in community settings. By providing an accurate 8-item alternative to lengthier clinical interviews, the ABQ facilitates early risk identification which can help at-risk youth receive timely support.

Introduction

Antisocial behaviour (AB) encompasses actions violating social norms, legal rules, or the rights of others, varying in frequency and developmental significance from early childhood (e.g. temperament) to adolescence (e.g. truancy) and into young adulthood (offending) (American Psychiatric Association [APA], 2013; Frick & Viding, 2009; Moffitt, 1993). Distinguishing adolescent behaviours in minors (<18 years) versus emerging adults (Arnett, 2000; Sawyer et al., 2018) is vital, as conduct problems often shift from overt school defiance to covert antisocial behaviours like property damage and theft which carry greater legal and clinical weight once individuals reach the age of majority (Moffitt, 1993; Farrington, 2003). Accordingly, our analyses used dataset that extended to age 24 to capture continuity of antisocial behaviour across adolescence and emerging adulthood, a developmental period increasingly recognised as part of a prolonged transition to adult psychosocial functioning (Arnett, 2000; Sawyer et al., 2018). Conduct Disorder (CD) is a psychiatric diagnosis for minors (< 18 years) requiring ≥ 3 of 15 antisocial symptoms that cause functional impairment, and in adults, such antisocial symptoms contribute to a diagnosis of antisocial personality disorder (APA, 2013). While CD is the categorical clinical “endpoint”, AB represents the entire dimensional continuum (Frick & Viding, 2009). Theoretically, this continuum is grounded in the externalizing spectrum framework positing that diverse antisocial acts manifest a singular, latent propensity for disinhibition (Krueger et al., 2005). Precise measurement of AB is critical for distinguishing normative adolescent defiance from early markers of persistent conduct pathology characterized by intentional harm or systemic rule-breaking (Frick & Viding, 2009).

Abundant validated tools for externalising behaviours exist, but many lack the diagnostic depth required to capture the complex and developmentally changing antisocial phenotype. Measure selection is hampered by labelling inconsistencies, where identical behaviours are

labelled differently or distinct acts share the same name (Marsh, 1994). Many broadband externalising scales consequently lack essential CD components (Waschbusch & Walters, 2006; Panchal et al., in preparation). For instance, the Strengths and Difficulties Questionnaire (SDQ) Conduct Problems scale prioritises aggression but excludes property destruction or deceit (Goodman, 1997). Similarly, broad-band measures designed to assess multiple domains pf psychopathology, including internalising and externalising difficulties such as the Youth Self-Report (YSR; Achenbach & Rescorla, 2001) often miss covert theft, unless extended versions with substantial respondent burden are used (e.g. 338 items; Achenbach, Dumenci, & Rescorla, 2003).

To map the measurement landscape, we conducted a systematic taxonomic audit of 2,378 records of adolescent measures assessing externalising and behavioural difficulties, many of which were not originally designed to assess Conduct Disorder specifically (Panchal et al., in preparation). This identified 47 instruments containing conduct-problem content, of which 27 (57%) were “Strongly Aligned” with all four DSM-5 CD domains and 21 (45%) also demonstrated evidence of structural validity, internal consistency and measurement invariance (Panchal et al., in preparation). Analysis of these 21 measures revealed a critical inverse relationship between diagnostic rigour and respondent burden. Among tools with comprehensive DSM-5 CD alignment and psychometric stability, none had 10 or fewer items; five had 20 or less, while the remainder required 50-300+ (Panchal et al., in preparation). Within this review, a subset of "Strongly Aligned" tools emerged as gold-standard diagnostic surrogates–notably the ASEBA (CBCL, YSR, and TRF; 338 items; Achenbach & Rescorla, 2001) and the BASC-3/BASC-BESS (370-516 items; Reynolds & Kamphaus, 2015). While offering exceptional precision (pooled $\alpha = .89$), these exceed 300 items, typically require paid licensing, and involve high respondent burden that can hinder large-scale data collection.

Beyond length, systemic barriers limit accessibility. Some measures require administration by trained clinicians, whereas others rely exclusively on parent or teacher observations (Panchal et al, in prep). While stable, such external informant measures may miss private, subtle, or peer-based (or other contexts) antisocial acts best captured via youth self-report (Biderman & Reiss, 1967; De Los Reyes et al., 2015). Conversely, some specialised measures are primarily designed for high-risk or forensic populations (Luo et al., 2021). Consequently, a void remains: no current instrument simultaneously offers ultra-brief administration (≤ 10 items), comprehensive DSM-5 CD alignment, with robust psychometric stability within an open-source, youth self-report format. This lack of a practical screening tool hinders the ability to reliably capture the full antisocial phenotypic across diverse educational, clinical and community settings.

The Antisocial Behaviour Questionnaire (ABQ) addresses these requirements by providing a bridge between conceptual alignment and practical feasibility. Originally developed for the ROOTS study (Goodyer et al., 2010) and subsequently implemented in the NeuroScience in Psychiatry Network (NSPN) cohort (Kiddle et al., 2018), the ABQ is a short, behaviourally specific index of concrete antisocial acts comprehensively aligned to DSM-5 CD domains. ABQ is designed to provide a unified index of antisocial behaviours for efficient clinical screening (Caspi et al., 2014; Krueger et al., 2005). As an open-access tool, the ABQ captures low-frequency but high-severity behaviours while remaining efficient enough for longitudinal tracking and routine practice. Establishing whether the ABQ functions as a psychometrically stable, unidimensional index with robust measurement invariance and a relevant clinical cut-off is therefore critical for its utility as a concise measure in both research and clinical settings.

The present study aims to evaluate the psychometric robustness of the ABQ using data from two independent UK community cohorts as follows: (1) establishing dimensionality via

EFA/CFA and assessing internal consistency, (2) testing measurement invariance across gender, age bands (14-24), and longitudinal waves, (3) assessing construct validity against established measures of aggression and psychopathic traits, (4) testing predictive validity by evaluating the ABQ's capacity to predict future substance misuse, and (5) assessing diagnostic utility by determining optimal clinical thresholds against "gold-standard" structured diagnostic interviews. Establishing these properties provides an evidence-based foundation for the ABQ as a high-precision, low-burden diagnostic screener, which will enable the ABQ to efficiently stratify risk across community-based adolescents.

Method

Participants and Data Sources

Data were drawn from two longitudinal community cohorts focused on developmental mental health from mid-adolescence into early adulthood (ages 14-24): the ROOTS project, which investigated risk patterns and processes for psychopathology in adolescence (Goodyer et al., 2010), and the NeuroScience in Psychiatry Network (NSPN 2400) cohort, designed to characterise normative and abnormal developmental change across the post-pubertal decade (Kiddle et al., 2018). At the time of analysis, ROOTS and NSPN were the only datasets known to have administered the ABQ, which determined their selection for this evaluation. The ABQ formed part of the Home Questionnaire Pack at all waves for both cohorts (Table S1).

ROOTS. Participants were recruited via 18 secondary schools in Cambridgeshire and Suffolk; 3,762 students were invited and 1,238 consented at baseline (M= 14.5 years, SD = 0.5; 55% female). Data collection for the three primary adolescent waves (baseline, Wave 2, and Wave 3) spanned from 2005 to 2010. Assessments occurred at baseline and again approximately 18 and 36 months later. Baseline and follow-up included postal questionnaire packs (completed at home) and semi-structured diagnostic interviews (Goodyer et al., 2010).

NSPN 2400. Participants aged 14–24 years ($N \approx$ 2,400) were recruited through GP practices, schools, and via community outreach. The NSPN sample included participants from both Cambridgeshire and major urban centres, such as Greater London, and thus was more socio-demographically diverse than the ROOTS cohort. Data collection for the corresponding waves (baseline to Wave 3) took place between 2012 and 2017. Sampling was age and sex stratified by age band (14–15, 16–17, 18–19, 20–21, 22–24) to ensure adequate representation across the transition from adolescence to young adulthood; reassessments

occurred at approximately 12 and 24 months (Waves 2-3), enabling analyses of stability and change (Kiddle et al., 2018).

Analytic samples and strategy. Analytic stages and corresponding datasets are summarised in Table S2. We used the ROOTS baseline sample (N = 1,153; mean age = 14.5 years; SD = 0.5) for the initial structural evaluation as this cohort provided a large, developmentally homogenous sample for establishing and cross validating the core factor structure. Furthermore, preliminary data inspection revealed that the ROOTS baseline exhibited higher endorsement rates for key antisocial acts compared to the NSPN cohort. This resulted in higher item-level variances, providing a more robust foundation for the primary EFA. This sample size was sufficient to allow for a random split into two independent halves (n > 500): split half1 (n = 577) was used for exploratory factor analysis (EFA) and split half2 (n = 576) for confirmatory factor analysis (CFA). The full ROOTS baseline sample was subsequently used for measurement invariance testing. Once the structure was established in ROOTS, it was cross-validated in the 14-15-year-old NSPN baseline cohort. Finally, longitudinal and multi-group analyses utilized all available cases in NSPN under age 18 ($N \approx$ 2,400 at baseline) with specific sample sizes (n) reported alongside results to account for expected attrition.

Measures

Antisocial Behaviours Questionnaire (ABQ). The Antisocial Behaviours Questionnaire (ABQ) was developed as a concise (11 items), clinically grounded, DSM CD aligned measure of antisocial acts for use in a large-scale longitudinal study – ROOTS (Goodyer et al., 2010). Originally called the “Behaviour Checklist”, items were developed in 2004 by a working group (led by Ian Goodyer) of clinical and research psychiatrists and psychometricians specialising in antisocial behaviour development in children and adolescents. The original 11-item pool was designed to capture the core behavioural topography of the DSM Conduct

Disorder criteria across four major domains: aggression to people and animals, destruction of property, deceitfulness or theft, and serious violations of rules (Table 1). Selection prioritised items indexing a continuum of behaviours from normative adolescent experimentation to severe psychopathology. Items were originally rated on a four-point Likert scale (0 = never; 3 = always) to capture frequency and intensity. To support longitudinal and cross-cohort comparisons, the measure retained identical wording and item ordering across all study administrations.

Comparator Measures. *Convergent* validity was examined using constructs theoretically proximal to antisocial behaviour (Table S3): the Antisocial Process Screening Device (APSD; callous–unemotional, narcissism, impulsivity; (Frick & Hare, 2001)), the Inventory of Callous–Unemotional Traits (ICU; callous, uncaring, unemotional; (Essau et al., 2006; Kimonis et al., 2008)), the Barratt Impulsiveness Scale—11 (BIS-11; attentional, motor, non-planning impulsivity; (Patton et al., 1995)), and the Child and Adolescent Dispositions Scale (CADS; Daring, Negative Emotionality; (Lahey et al., 2008)). *Divergent* validity was assessed via prosocial tendencies using the Strengths and Difficulties Questionnaire (SDQ Prosocial; Goodman, 1997, 2001); and CADS Prosociality). *Predictive* validity was assessed via Drug and Alcohol Substance Involvement (DASI; Kiddle et al., 2018). *Diagnostic* validity was established against clinical assessment using K-SADS-PL interviews in ROOTS to identify conduct disorder and a broader composite of any externalising disorder (Kaufman et al., 1997).

Data Preparation and Missingness

To accommodate the low-frequency nature of the antisocial acts measured and ensure model stability, all original ABQ items were binarised (0 = never; 1 = any endorsement). This addressed distributional requirements for latent variable modelling with sparse categorical data (Brown, 2015). Initial inspection revealed that three items, *extortion or robbery* (ABQ8),

breaking into property (ABQ9), and *animal cruelty* (ABQ11), were endorsed <2% in both cohorts. These eight items were summed to create total (range 0–8) scores, with higher totals indicating greater antisocial involvement.

Comparator measures were scored according to established study protocols for the ROOTS and NSPN cohorts. Total scores were calculated pro-rata where at least 85% of items within a measure were present; otherwise, the score was treated as missing. For all subsequent validation and diagnostic utility analyses, correlations and ROC curves were computed using pairwise deletion for comparator measures after ensuring listwise-complete data for the primary 8-item ABQ set to maximize the use of available data (Little & Rubin, 2019). Consistent with prior work in these cohorts where missing data over time was predicted by a wealth of observed variables, longitudinal missing data were presumed to be Missing at Random (MAR; e.g., ROOTS: Neufeld et al., 2017; NSPN: Polek et al, 2020). While maximum likelihood and multiple imputation represent recommended MAR-consistent approach (Little & Rubin, 2019; Schafer & Graham, 2002), analysis-specific listwise deletion to maintain congruence between the implied model and the observed covariance structure under WLSMV estimation. This approach is supported for longitudinal measurement invariance (MI) testing with categorical indicators (Chen et al, 2020).

Attrition remained within simulation-supported bounds for unbiased scalar invariance estimation. In the ROOTS cohort, analytic samples were n = 1,154 (Baseline), n = 874 (Wave 2; 24.3% missing), and n = 999 (Wave 3; 13.4% missing). In NSPN, samples were n = 2,378 (Baseline), n = 1,680 (Wave 2; 29.4% missing), and n = 1,127 (Wave 3; 52.6% missing). Across measurement invariance analyses, most comparison groups exceeded n ≥ 500, with only the NSPN Wave 3 male subgroup approaching this threshold (n = 441). Simulation studies indicate that scalar invariance bias under WLSMV estimation remains acceptably low when group sizes exceed approximately 500 and missingness remains below ~50% (Chen et

al., 2020); therefore, analytic conditions were considered adequate or closely approximated recommended parameters for unbiased estimation.

Statistical Analyses

All analyses were conducted in R version 4.0.4 (R Core Team, 2021). The psych package (Revelle, 2020) was used for tetrachoric correlation matrices, EFA, and estimating reliability. The lavaan package (Rosseel, 2012) was used for CFA and MI testing, utilizing weighted least squares mean and variance adjusted (WLSMV) estimation with θ-parameterization. Diagnostic utility was quantified through receiver operating characteristic (ROC) estimation using the pROC package (Robin et al., 2011). To facilitate transparency and replication, all analysis scripts, output files, and table shells are available from the authors upon request.

Factor Structure (EFA/CFA)

EFA. Exploratory Factor Analysis was conducted on the ROOTS baseline split half1 (*n* = 577) using a tetrachoric correlation matrix. Factorability was confirmed via Bartlett's test of Sphericity (Bartlett, 1950 and the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy (Kaiser, 1970). Factor retention was determined through parallel analysis (Horn, 1965), scree plot inspection, and fit-index/residual change (change in Root Mean Square Error of Approximation - ΔRMSEA and residual reduction) when comparing *k* versus *k*+1 solutions (Finch, 2019). Extraction utilized minimum residual (minres) with oblimin rotation. In accordance with the externalizing spectrum framework and the theoretical unidimensionality detailed in the introduction, we expected a dominant first factor; however, a two-factor alternative was examined for completeness.

CFA. Confirmatory Factor Analysis was performed on the ROOTS split half2 (n = 576) to cross-validate findings using WLSMV estimation (Brown, 2015; Flora & Curran, 2004). Model fit was evaluated using χ^2, Comparative Fit Index (CFI), Tucker–Lewis Index (TLI),

Root Mean Square Error of Approximation (RMSEA) with 95% confidence intervals, and Standardized Root Mean Square Residual (SRMR). Conventional benchmarks for excellent fit were applied (*CFI/TLI* ≥.95, RMSEA ≤ .06, and SRMR ≤ .08), with χ^2 interpreted cautiously due to sample size sensitivity (Hu & Bentler, 1999; Brown, 2015). To test the hypothesized unidimensional structure, the model excluded cross-loadings or correlated residuals, ensuring a parsimonious fit aligned with externalising theory (Krueger et al., 2005). Modification indices were consulted solely for model diagnostic purposes (Brown, 2015). We examined standardized loadings and thresholds, utilising 1,000-sample bootstrap confidence intervals to ensure parameter stability (Efron & Tibshirani, 1994; Nevitt & Hancock, 2001). Finally, generalizability was tested by fitting the same model to the NSPN 14–15-year-old baseline subsample using identical procedures (Brown, 2015; Rosseel, 2012).

To evaluate whether summed ABQ scores were psychometrically justified, a tau-equivalent CFA model constraining all factor loadings to equality was compared with the freely estimated congeneric model in the CFA validation samples (ROOTS split-half confirmatory sample: N = 576; NSPN baseline 14–15 age band: N = 464). In both cohorts, imposing equal loadings did not significantly worsen model fit (ROOTS: $\Delta\chi^2(7) = 11.72$, $p = .11$; NSPN: $\Delta\chi^2(7) = 7.70$, $p = .36$), and changes in incremental fit indices were negligible (ΔCFI ≤ .015; ΔRMSEA ≈ 0). These findings indicate that the assumption of tau-equivalence was not violated, supporting interpretation of ABQ summed scores alongside factor-based reliability estimates (Schlechter et al., 2022).

Measurement Invariance

Measurement invariance (MI) was evaluated across three dimensions to ensure structural stability: (1) gender (males vs females) in both cohorts across Waves 1–3, examined both within individual waves (cross-sectional multi-group models) and within a single longitudinal multi-group framework spanning Waves 1–3; (2) age across NSPN developmental age bands

at baseline (14–15, 16–17, 18–19, 20–21, and 22–24 years); and (3) gender invariance examined within consolidated developmental age bands (14–17 and 18–24 years) at NSPN baseline to evaluate gender equivalence within developmental strata. Analyses were conducted using analytic subsamples with sufficient non-missing data at each wave (see Data Preparation and Missingness), with subgroup sample sizes reported in Tables S10–S11. All multi-group confirmatory factor analyses were estimated using weighted least squares mean- and variance-adjusted (WLSMV) estimation with binary indicators. Following recommendations for categorical indicators, invariance testing proceeded sequentially from configural (same factor structure) to scalar (equal loadings and thresholds) to strict (equal residual variances), bypassing the metric stage (Neufeld et al., 2024). Model comparisons prioritised incremental fit indices to reduce sensitivity of χ^2 statistics to large samples. Measurement invariance was evaluated through comparisons of nested models using ΔCFI < .010 and ΔRMSEA < .007 criteria (Chen, 2007; Sass et al., 2014). Because ΔRMSEA may be unstable when subgroup sizes approach $n \approx 500$, this index was interpreted cautiously in smaller groups. Invariance was considered supported when at least one incremental fit index met criteria and overall model fit remained acceptable. Scaled χ^2 difference tests (Satorra & Bentler, 2001) were reported for completeness but were not used as primary decision criteria. Overall model fit followed guidelines synthesised by Brown (2015): CFI/TLI $\geq$ .95 and RMSEA $\leq$ .06 indicated reasonably good fit, whereas CFI/TLI $\geq$ .90 reflected acceptable fit (Bentler, 1990). RMSEA values between .08–.10 indicated mediocre fit, and RMSEA $\geq$ .10 indicated poor fit (MacCallum et al., 1996; Brown, 2015). After establishing invariance at the highest supported level, latent mean differences were estimated by fixing reference group means to zero (Meredith, 1993; Putnick & Bornstein, 2016). Because invariance cut-offs are derived primarily from ordered-categorical simulations, findings were interpreted cautiously for binary indicators (Sass et al., 2014).

Reliability

Following the CFA to confirm model fit, reliability estimates were calculated using the full ROOTS baseline sample (N = 1,153). This maximizes statistical power and ensures that reported reliability estimates are representative of the full study population (Cudeck & Browne, 1983; Harrell, 2017). Internal consistency was evaluated using Cronbach's α and McDonald's ω (Cronbach, 1951; McDonald, 1999; Zinbarg et al., 2005). While α is standard, it assumes tau-equivalence (equal factor loadings). This assumption was formally evaluated within the CFA framework and was not violated. We nevertheless prioritized McDonald's ω, a factor-based estimator that accommodates varying loadings. To account for the binary nature of the ABQ, ω was derived from a tetrachoric correlation matrix, providing a more accurate estimate of the underlying latent construct's reliability (McDonald, 1999; Zinbarg et al., 2005).

Construct Validity

Construct validity was evaluated by correlating ABQ total scores with established comparator measures using Pearson's r with two-tailed tests and Fisher's z 95% CIs. Analyses were primarily cross-sectional at baseline in the ROOTS study, while the NSPN cohort provided these external measures across Waves 1–3, enabling longitudinal replication of the validity patterns.

Our primary hypotheses focused on establishing convergent validity through anticipated positive associations with measures of callous-unemotional traits (ICU; Essau et al., 2006; Kimonis et al., 2008), psychopathic tendencies (APSD; Frick & Hare, 2001), impulsivity (BIS-11; Patton et al., 1995), and the negative emotionality and daring subscales of the CADS (Lahey et al., 2008). Divergent validity was hypothesized via negative associations with prosocial behaviour scales from both the SDQ (Goodman, 2001) and the CADS (Lahey et al., 2008).

Predictive validity was assessed by correlating baseline ABQ scores with Wave 3 substance use outcomes derived from the Drugs, Alcohol, and Self-Injury questionnaire (St Clair et al, 2017). Given the non-normal distribution of substance use frequency data, Spearman's ρ with Fisher's z confidence intervals (CIs) was utilized as a robustness check. Substance use items were categorized into legal substances (e.g., frequency of alcohol and tobacco use) and illegal drug classes (e.g., cannabis and other illicit substances). For rare drug classes with low endorsement (e.g., cocaine or hallucinogens), binary indicators were utilized to ensure stable estimates during predictive modelling.

Diagnostic Utility (ROC)

We assessed the ABQ's ability to discriminate between participants with and without a Conduct Disorder diagnosis using the Kiddie-SADS-Present and Lifetime Version (K-SADS-PL DSM-IV; Kaufman et al., 1997), using receiver operating characteristic (ROC) curves and DeLong 95% confidence intervals (CIs) for the Area Under the Curve (AUC; DeLong et al., 1988; Robin et al., 2011). Clinical utility was assessed through both concurrent validity (using ABQ and K-SADS data collected within the same wave at Baseline and Wave 3) and predictive validity (Baseline ABQ scores predicting Wave 3 diagnostic status). Analyses utilized the binarized ABQ total score, with optimal clinical thresholds identified via Youden's index (Youden, 1950). Because predictive values are prevalence-dependent, we derived positive predictive values (PPV) and negative predictive values (NPV) at plausible community base rates to aid screening interpretation (Altman & Bland, 1994). Given the low prevalence of certain diagnostic outcomes, we interpreted AUC and standard errors (SE) with caution, following established guidelines for diagnostic accuracy in community samples (Florkowski, 2008; Hajian-Tilaki, 2013). For clinical context, AUC interpretation followed conventional benchmarks (e.g., .70–.80 as acceptable, .80–.90 as excellent; Youden, 1950).

Ethical Considerations

Original data collection for the ROOTS and NSPN cohorts was approved by UK National Health Service (NHS) Research Ethics Committees. This secondary analysis was approved by the University of East Anglia Faculty of Medicine and Health Research Ethics Committee (Ref: ETH2425-2404). All participants provided written informed consent; parental/guardian consent was obtained for those under 16. Detailed procedures are available in the primary cohort reports (Goodyer et al., 2010; Kiddle et al., 2018).

Results

Descriptive and Preparation Checks

Demographics (Table S1), including ethnicity and socioeconomic status (SES), and item-level endorsement patterns (Table S4) were consistent across both ROOTS and NSPN and typical for community samples. Common behaviours like rule breaking/disobedience (ABQ1; 38.1%–43.0%) and lying/cheating (ABQ6; 22.2%–28.9%) were moderately endorsed, whereas more severe acts like extortion or robbery (ABQ8; 0.4%), breaking into property (ABQ9; 0.6%), and animal cruelty (ABQ11; 1.6%) fell below the 2% threshold at the baseline. These items were excluded due to extremely low endorsement (<2%), which produced unstable parameter estimates and convergence difficulties in preliminary categorical CFA models (Brown, 2015; Flora & Curran, 2004), as evidenced by convergence issues when models were explored including these items. Post-binarisation of the 8-item set, χ^2 tests confirmed equivalent endorsement across the randomly split ROOTS halves ($p > .05$ for all retained items) and the NSPN comparison group.

Data checks confirmed factorability. Kaiser-Meyer-Olkin (KMO) values were adequate across all primary analytic sets (.69 for ROOTS split half1; .80 for ROOTS split half2; .69 for NSPN 14–15; Kaiser, 1970). Bartlett's test of sphericity was significant in all samples ($\chi^2(28)$ range: 1564.06 to 4353.66, $p < .001$; Bartlett, 1950), indicating sufficient common variance for factor modelling. Tetrachoric correlation matrices are provided in the Supplementary Materials (Tables S5, S6, and S7, respectively).

Factor Structure

Exploratory Factor Analysis (ROOTS split half1). Horn's Parallel Analysis (Figure S1) indicated a one factor solution, with only one observed eigenvalue exceeding those from random data. The scree plot (Figure S2) corroborated this, showing a clear "elbow" after the

first factor. The ΔRMSEA approach showed negligible improvement between 1-factor (RMSEA = 0.259) and 2-factor (RMSEA = 0.258) models, also supporting a parsimonious unidimensional structure (Finch, 2020). The 1-factor solution explained 56% of common variance; standardised loadings ranged from .66 to .83, with communalities ≥ .44 (Table S8). Results indicate high internal consistency (off-diagonal fit = .98, RMSR = .08).

Confirmatory Factor Analysis (ROOTS split half2; NSPN baseline ages 14-15). The unidimensional ABQ model demonstrated excellent fit in both cohorts (Table 2A). In ROOTS split half2, indices were robust without post-hoc modifications: $\chi^2(20) = 22.09$, $p = .34$; CFI = .999; TLI = .998; SRMR = 0.055; RMSEA = .013, with 90% CI [.000, .039]. Standardised loadings were large and statistically significant (.70 – .94) (Table 2B). The model also fit well in the NSPN age 14–15 subsample: $\chi^2(20) = 15.10$, $p = .77$; CFI = 1.000; TLI = 1.020; RMSEA = .000; SRMR = .103. The SRMR was significantly higher in NSPN ages 14–15 (.103, 95% CI [.089, .117]) compared to ROOTS split half2 (.055, 95% CI [.048, .062], but still within acceptable range (≤ .10 for categorical models with high sparsity and low item endorsement, Maydeu-Olivares, 2017; Shi et al., 2020). Given that incremental and absolute fit excellent (CFI = 1.000; TLI = 1.020; RMSEA = .000) this was not interpreted as evidence of model misspecification. Bootstrap confidence intervals based on 1,000 resamples were narrow in both samples, indicating stable parameter estimation (Table S9). Factor loadings in NSPN (.48–.85) were broadly comparable to ROOTS (.70-.94), although items reflecting rule-breaking/disobedience, theft, and property damage showed modestly weaker associations with the latent factor in the NSPN subsample, consistent with higher residual variances (Table S9). Competing specifications were examined; neither a two-factor model ($\chi^2(19) = 21.50$; ΔCFI < .001) nor a bifactor specification ($\chi^2(12) = 14.20$; ΔRMSEA < .005) offered a significant improvement in fit or a clear theoretical advantage once model complexity was

considered. Given that all primary indices of incremental and absolute fit remained excellent, the parsimonious single-factor solution was retained for both cohorts.

Measurement invariance

Gender Invariance (ROOTS and NSPN, Baseline to Wave 3): Measurement invariance was first evaluated across gender to establish whether ABQ scores could be meaningfully compared between males and females. Gender invariance was first examined separately at each wave in both cohorts (Table S10A). Across ROOTS (ages ~14.5–17.5 years) and NSPN (ages 14–26 years), scalar invariance was supported at all waves, with minimal changes in incremental fit indices (all ΔCFI ≤ .008; ΔRMSEA ≤ .012), indicating comparable measurement properties across males and females. Although absolute fit declined at ROOTS Wave 3, incremental fit criteria remained stable, consistent with increased sparsity and attrition at later waves. To account for potential bias associated with wave-specific attrition, gender invariance was re-evaluated using a single longitudinal multi-group model spanning Waves 1–3 within each cohort (Table S10B). In NSPN, the "carrying a weapon" item showed extremely low endorsement (<2%), which prevented stable threshold estimation in longitudinal WLSMV models; this item was therefore excluded from NSPN longitudinal invariance analyses but retained in cross-sectional models. In ROOTS, imposing scalar constraints did not worsen fit relative to the configural model (ΔCFI = +.003; ΔRMSEA = −.001), and strict invariance was retained. The same pattern was observed in NSPN (N = 2,403), where configural, scalar, and strict models showed virtually identical fit (ΔCFI ≤ .001; ΔRMSEA = .000), supporting longitudinal gender invariance. Together, these findings indicate that ABQ measurement properties are stable across gender both cross-sectionally and longitudinally, supporting meaningful comparison of latent means across males and females.

Age-band invariance (NSPN baseline, ages 14–24): Having established gender equivalence, measurement stability across developmental stages was examined using NSPN

baseline data. Configural invariance was supported across all five NSPN age bands ($\chi^2(100) = 113.56$, $p = .167$; CFI = .997; RMSEA = .017; Table S11A), indicating a consistent factor structure across developmental stages. However, scalar and strict models showed reductions in incremental fit (scalar: ΔCFI = −.010, ΔRMSEA = +.013; strict: ΔCFI = −.033, ΔRMSEA = +.020), suggesting departures from full invariance. Examination of item endorsement patterns (Table S4) indicated extremely low prevalence of several low-frequency behaviours in the youngest group (14–15 years), particularly property damage and running away (≈2–3% endorsement), consistent with unstable threshold estimation under binary CFA. To determine whether this reflected developmental sparsity rather than structural differences in the latent construct, a sensitivity analysis excluding the 14–15 age band was conducted. In the reduced four-group model (ages 16–24), configural, scalar, and strict invariance were supported (Table S11A), with minimal changes in fit (ΔCFI ≤ −.003; ΔRMSEA ≤ +.003) and good overall model fit (CFI ≥ .988; RMSEA ≤ .027), indicating stable measurement properties from mid-adolescence into young adulthood. Latent mean comparisons (reference = 14–15 years) showed higher antisocial behaviour in the 16–17 age band ($\Delta M = +0.366$, $p = .036$), followed by significantly lower latent means in young adulthood (18–19 years: $\Delta M = -0.605$, $p = .011$; 20–21 years: $\Delta M = -0.678$, $p = .006$), while the 22–24 group did not differ significantly from the reference group ($\Delta M = -0.079$, $p = .614$).

Developmental gender invariance (age-stratified multi-group models; NSPN baseline): Finally, gender invariance was evaluated within consolidated developmental age bands (14–17 and 18–24 years) to determine whether gender equivalence was maintained within distinct developmental stages (Table 3). Configural, scalar, and strict invariance were supported in both age bands, with minimal changes in incremental fit indices (ΔCFI ≤ .019; ΔRMSEA ≤ .012) and acceptable overall model fit. These findings indicate that the ABQ functions equivalently for males and females within both mid-adolescent and young adult

developmental periods. Latent mean comparisons showed higher antisocial behaviour in males within the 14–17 age band (ΔM = −0.469, p = .015), whereas no gender difference was observed in the 18–24 group (ΔM = +0.028, p = .836). Together, these results support gender comparability of ABQ scores across developmental stages while suggesting that gender differences are most pronounced during mid-adolescence.

Reliability

In the full ROOTS baseline sample, the 8-item ABQ showed acceptable to excellent internal consistency (Table S12). McDonald's ω was 0.93, while the more conservative Cronbach's α was 0.73 (95% CI [0.71, 0.75]). This discrepancy suggests α is attenuated by the scale's binary items and uneven endorsement; ω provides a more robust estimate by relaxing tau-equivalence assumptions (McNeish, 2018). Scale cohesion was further supported by an average inter-item correlation of r = .28, and perfect unidimensionality (*ECV* = 1.00). Factor score adequacy was high (correlation with factor = .97), indicating factor scores are highly representative of the underlying construct. Item-level diagnostics (Table S13) confirmed uniformly positive item-total correlations and that no item deletion improved the overall α, supporting the retention of all eight items.

Diagnostic Utility (ROC)

Using K-SADS-PL clinical interviews as the gold-standard criterion (Kaufman et al., 1997), ROOTS baseline ABQ scores demonstrated excellent concurrent discrimination for Conduct Disorder (AUC = .854, 95% CI [.823, .886]) and acceptable discrimination for a broader category of any externalising disorder (AUC = .768, 95% CI [.731, .804]). The Youden index (Youden, 1950) identified a cut-off of ≥2 endorsed behaviours as the optimal balance between sensitivity and specificity (Sensitivity = .846; Specificity = .739). Higher thresholds (e.g., ≥3) increased specificity (> .869) but reduced sensitivity (.679), reflecting the expected trade-off

between false positives and false negatives. Accordingly, a threshold of ≥2 may be appropriate for research screening contexts requiring balanced classification, whereas higher thresholds may favour specificity when false positives must be minimised; however, optimal clinical thresholds should be confirmed in clinically ascertained samples. These findings indicate that ABQ thresholds function as probabilistic risk indicators rather than diagnostic decision rules. Threshold-specific operating characteristics are reported in Table S14, including positive and negative predictive values (PPV/NPV). Concurrent and longitudinal predictive models from Baseline to Wave 3 showed similar but attenuated discrimination. This attenuation likely reflects multiple factors, including developmental change, reduced outcome prevalence over time, and the influence of unmeasured intervening variables. Consistent with developmental analyses, ABQ latent means peaked in mid-adolescence (16–17 years) and declined thereafter (Table S11), suggesting that reduced predictive discrimination at follow-up partly reflects normative developmental decreases in antisocial behaviour rather than reduced measurement validity.

Construct Validity

As comparator measures in ROOTS were available only at baseline, longitudinal replication was examined in the NSPN cohort across Waves 1–3. As hypothesised, ABQ total scores demonstrated a consistent pattern of convergent validity across both cohorts and waves (Tables 5 and S15; Figure S3). The ABQ showed large positive associations with antisocial and psychopathy-related traits, specifically APSD Total scores (Frick & Hare, 2001) in both ROOTS (r = .53) and NSPN (Baseline r = .53; Wave 3 r = .52). Moderate associations were observed with hyperactivity (SDQ; r = .32) and impulsivity (BIS-11; r = .33–.41; Patton et al., 1995). Within the APSD, correlations were strongest for the Impulsivity (r = .39–.44) and Narcissism (r = .37–.40) dimensions. Similarly, ICU associations (ICU total r = .42–.44; Essau et al., 2006; Kimonis et al., 2008) followed expected patterns, with stronger relations

for Callous (r = .44–.47) and Uncaring (r = .34–.37) dimensions than for the Unemotional subscale (r = .16–.18). A smaller but significant association was observed with the CADS Daring subscale (r = .17–.22; Lahey et al., 2008). Divergent validity was supported by consistent inverse correlations with prosocial behaviour indices in both ROOTS (SDQ; r = −.28; Goodman, 2001) and NSPN (CADS; r = −.26–−.30; Lahey et al., 2008). Full construct validity coefficients across all measures, subscales, and waves are reported in Table S15. Predictive validity was assessed in ROOTS by correlating baseline ABQ scores with substance use outcomes approximately 36 months later (Wave 3; DASI; Kiddle et al., 2018). Baseline ABQ scores significantly predicted all substance use outcomes at follow-up, including total substance misuse (ρ = .151, 95% CI [.094, .207]) and illegal drug involvement (ρ = .220, 95% CI [.164, .274]; Table 5). Detailed prospective associations for all outcomes are provided in Table S15.

Discussion

The results across the ROOTS and NSPN cohorts demonstrate that the 8-item ABQ is a psychometrically robust, unidimensional measure of antisocial behaviour. The scale exhibited a stable factor structure, strong reliability, and consistent evidence of concurrent and predictive validity across the transition from adolescence to young adulthood. Multi-group analyses further supported measurement stability across gender and developmental stages, enabling meaningful comparison of antisocial behaviour across adolescence (Meredith, 1993; Putnick & Bornstein, 2016). Gender was measured using binary categories in both cohorts; future research should examine whether measurement properties remain equivalent across non-binary and gender-diverse populations. These findings suggest the ABQ captures a spectrum of antisocial behaviours ranging from relatively common rule-breaking acts to less frequent but more severe behaviours like damage to property or carrying a weapon within community samples. However, its performance in forensic or clinical populations remains to be directly evaluated (Edens et al., 2001).

The ABQ: Psychometric Foundation and Developmental Stability.

The 8-item ABQ demonstrated a robust unidimensional structure across both cohorts, explaining approximately 56% of common variance with excellent internal consistency. This level of coherence supports essential unidimensionality and justifies interpretation using a single total score (Reise et al., 2013; Bonifay & Cai, 2017), consistent with dimensional models of externalising behaviour that prioritise a general liability toward rule-breaking over behavioural sub-typing (Caspi et al., 2014; Krueger et al., 2005).

Measurement invariance analyses indicated scalar invariance across gender and developmental stages once developmental sparsity in the youngest age group was addressed. Sensitivity analyses excluding the 14–15 age band supported scalar and strict invariance

across ages 16–24, indicating that observed developmental differences are unlikely to reflect measurement artefact (Chen, 2007; Sass et al., 2014). This strengthens interpretation of the observed mid-adolescent elevation in antisocial behaviour as consistent with established developmental patterns described by the age–crime curve (Moffitt, 1993; Farrington, 2005).

Strong associations with antisocial traits, impulsivity, and callous–unemotional characteristics further support alignment with the externalising spectrum framework (Krueger et al., 2005; Caspi et al., 2014). Notably, stronger relations with callous and uncaring traits relative to unemotionality are consistent with theoretical accounts emphasising reduced concern for others' rights rather than emotional flattening as a central feature of antisocial behaviour (Frick et al., 2014; Kimonis et al., 2008).

Item-level findings nevertheless indicated developmental variability. Differences involving the skived item likely reflect contextual or linguistic influences rather than changes in the underlying construct. The term may be interpreted inconsistently across developmental or cultural contexts, and comparisons of individual item endorsement should therefore be interpreted cautiously, particularly when using binary indicators (Putnick & Bornstein, 2016).

Low endorsement rates for several high-threshold behaviours also highlight an important measurement boundary. In longitudinal NSPN analyses, extremely low prevalence of "carrying a weapon" limited stable estimation in longitudinal models, meaning invariance of this behaviour across time could not be fully evaluated. Sparse endorsement is a known challenge in categorical CFA when modelling low-base-rate behaviours (Brown, 2015; Sass et al., 2014) and likely reflects the community nature of the NSPN cohort rather than limitations of the broader construct. Consequently, while the ABQ performs well as a dimensional index of antisocial involvement, stability of very low-frequency behaviours requires further evaluation in higher-risk samples.

Overall, despite minor item-level variation, the consistency of the latent structure supports use of the ABQ total score as a defensible indicator for community-level risk stratification.

The ROC analyses further suggest that endorsing two or more behaviours provides a balanced screening threshold in community samples (Youden, 1950), whereas higher thresholds may prioritise specificity. However, optimal clinical cut-offs require validation in clinical or forensic populations before implementation in applied settings (Altman & Bland, 1994).

Strengths and Limitations

A major strength of this study is the dual-cohort design, allowing replication across independent UK samples collected using different recruitment strategies. Convergent findings across cohorts increase confidence that results are not sample-specific (Shrout & Rodgers, 2018). Additionally, alignment of items with DSM-5 conduct disorder domains ensures conceptual breadth often absent in brief screening measures (Haynes et al., 1995).

Several limitations warrant consideration. First, to maintain model stability, three items — extortion/robbery, breaking into property, and animal cruelty — were excluded during data preparation due to endorsement rates <2%. In addition, the “carrying a weapon” item showed extremely low endorsement in NSPN (<2%), which prevented stable threshold estimation in longitudinal categorical CFA models. This item was therefore excluded from NSPN longitudinal invariance analyses but retained in cross-sectional models and all other psychometric evaluations. This reflects sparse behavioural endorsement rather than conceptual limitations of the construct. As a result, the ABQ is best characterised as indexing a general spectrum of antisocial involvement rather than extreme forensic behaviour (Brown, 2015).

Second, findings may not generalise to younger children. Lower engagement in severe antisocial acts among younger adolescents likely reflects reduced opportunity, closer adult supervision, and developmental differences in peer exposure rather than measurement limitations alone (Edelbrock et al., 1986; Moffitt, 1993). Floor effects in early adolescence therefore remain an important consideration for future research.

Although overall sample sizes were large, statistical power was reduced within highly stratified subgroup analyses, particularly when simultaneously dividing samples by age and gender. Longitudinal interpretation should also consider selective attrition, although invariance results from longitudinal models suggest measurement stability was largely preserved (Putnick & Bornstein, 2016).

Finally, developmental theories guiding this work predate contemporary digital environments. Future adaptations could incorporate behaviours reflecting modern contexts (e.g., coordinated online harassment, digital fraud, or rule-violating online group activities) while remaining conceptually distinct from cyberbullying constructs. As with all self-report measures, social desirability bias remains possible; multi-informant designs would strengthen future validation efforts (De Los Reyes et al., 2015).

Clinical Implementation

The ABQ addresses a longstanding trade-off between diagnostic breadth and assessment burden. Whereas comprehensive instruments provide detailed assessment at the cost of length and licensing barriers, brief screening tools often omit key conduct-related behaviours. As a free, brief self-report instrument, the ABQ offers an efficient method for assessing antisocial behaviour within research and community contexts.

Future studies should evaluate performance and optimal thresholds within clinical and justice-involved samples, where behavioural prevalence and decision thresholds may differ

substantially. With further validation, the ABQ may support early identification and monitoring of antisocial behaviour across developmental transitions while maintaining minimal respondent burden.

References

Achenbach, T. M., Dumenci, L., & Rescorla, L. A. (2003). DSM-oriented and empirically based approaches to predicting psychopathology from youths' self-reports of problems. *Journal of Child Psychology and Psychiatry*, *44*(6), 828–840. https://doi.org/fpxmws

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

Altman, D. G., & Bland, J. M. (1994). Diagnostic tests 2: Predictive values. *BMJ*, *309*(6947), 102. https://doi.org/p8n

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). https://doi.org/gq9666

Arnett, J. J. (2000). Emerging adulthood: A theory of development from the late teens through the twenties. American Psychologist, 55(5), 469–480. https://doi.org/10.1037/0003-066X.55.5.469

Asparouhov, T., & Muthén, B. (2018). *SRMR in Mplus*. Mplus Technical Report. http://www.statmodel.com/download/SRMR2.pdf

Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Psychology (Statistical Section)*, *3*(2), 77–85. https://doi.org/d4n8z2

Biderman, A. D., & Reiss Jr, A. J. (1967). On exploring the "dark figure" of crime. *The Annals of the American Academy of Political and Social Science*, *374*(1), 1-15.

Block, J. (1995). A contrarian view of the five-factor approach to personality description. *Psychological Bulletin*, *117*(2), 187.

Bonifay, W., & Cai, L. (2017). On the complexity of item response theory models. *Multivariate Behavioral Research*, *52*(4), 465–484. https://doi.org/gbpqj6

Bons, D., van den Broek, E., Scheepers, F., Herpers, P., Rommelse, N., & Buitelaar, J. K. (2013). Motor, emotional, and cognitive empathy in children and adolescents with autism spectrum disorder and conduct disorder. *Journal of Abnormal Child Psychology*, *41*, 425-443.

Bornstein, M. H. (2011). *Handbook of cultural developmental science*. Psychology Press.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). Guilford.

Caspi, A., Houts, R. M., Belsky, D. W., Goldman-Mellor, S. J., Harrington, H., Israel, S., Meier, M. H., Ramrakha, S., Shalev, I., Poulton, R., & Moffitt, T. E. (2014). The p factor: One general psychopathology factor in the structure of psychiatric disorders? *Clinical Psychological Science*, *2*(2), 119–137. https://doi.org/v6j

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, *14*(3), 464–504. https://doi.org/d5m36k

Cicchetti, D. (2016). *Developmental Psychopathology*. John Wiley & Sons.

Côté, S. M., Orri, M., Tremblay, R. E., & Doyle, O. (2017). Psychometric properties of the Social Behavior Questionnaire in a sample of preschool children. *Journal of Abnormal Child Psychology*, *45*(7), 1361–1371. https://doi.org/gbks6m

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/dmms88

Cudeck, R., & Browne, M. W. (1983). Cross-validation of covariance structures. *Multivariate Behavioral Research*, *18*(2), 147–167. https://doi.org/fnkvw8

Dawson, A., & Sullivan, S. (2017). Callous-unemotional traits in the psychosis-proneness spectrum. *Psychiatry Research*, *258*, 205-212.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated ROC curves: A nonparametric approach. *Biometrics*, *44*(3), 837–845. https://doi.org/fpv2sz

Edelbrock, C., Costello, A. J., Dulcan, M. K., Kalas, R., & Conover, N. C. (1986). Age differences in reliability of the psychiatric interview of the child. *Child Development*, *57*(1), 165–175. https://doi.org/cq7jbm

Edens, J. F., Marcus, D. K., Lilienfeld, S. O., & Poythress, N. G. (2001). Psychopathic, not psychopathy: Taxometric evidence for the continuous structure of psychopathy. *Journal of Abnormal Psychology*, *110*(4), 670–681. https://doi.org/cp2d82

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC Press.

Essau, C. A., Sasagawa, S., & Frick, P. J. (2006). Callous-unemotional traits in a community sample of adolescents. *Assessment*, *13*(4), 454–469. https://doi.org/dpctd8

Finch, W. H. (2019). *Exploratory factor analysis*. SAGE Publications.

Finch, W. H. (2020). Using fit indices to determine the number of factors in EFA of inventory data. *Structural Equation Modeling*, *27*(2), 218–227. https://doi.org/ghm549

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for CFA with ordinal data. *Psychological Methods*, *9*(4), 466–491. https://doi.org/c77zdf

Florkowski, C. M. (2008). Sensitivity, specificity, ROC curves and likelihood ratios. *The Clinical Biochemist Reviews*, *29*(Suppl 1), S83–S87.

Frick, P. J., & Hare, R. D. (2001). *Antisocial Process Screening Device: Technical manual*. Multi-Health Systems.

Frick, P. J., Ray, J. V., Thornton, L. C., & Kahn, R. E. (2014). Can callous-unemotional traits enhance the understanding, diagnosis, and treatment of serious conduct problems? *Psychological Bulletin*, *140*(1), 1–57. https://doi.org/f5nkmq

Frick, P. J., & Viding, E. (2009). Antisocial behavior from a developmental psychopathology perspective. *Development and Psychopathology*, *21*(4), 1111–1131. https://doi.org/b6mv6t

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. *Journal of Child Psychology and Psychiatry*, *38*(5), 581–586. https://doi.org/dbnkbq

Goodman, R. (2001). Psychometric properties of the SDQ. *Journal of the American Academy of Child & Adolescent Psychiatry*, *40*(11), 1337–1345. https://doi.org/fshdfv

Goodyer, I. M., Bacon, A., Ban, M., Croudace, T., & Herbert, J. (2010). Serotonin transporter genotype, life histories and depressive symptoms in late adolescence. *The British Journal of Psychiatry*, *197*(3), 180–188. https://doi.org/dfxpgj

Goodyer, I. M., Croudace, T., Dunn, V., Herbert, J., & Jones, P. B. (2010). Cohort profile: The ROOTS project. *International Journal of Epidemiology*, *39*(2), 361–369. https://doi.org/dvjtcc

Hajian-Tilaki, K. (2013). ROC curve analysis for medical diagnostic test evaluation. *Caspian Journal of Internal Medicine*, *4*(2), 627–635.

Harrell, F. E., Jr. (2017). *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis*. Springer.

Haynes, S. N., Richard, D. C., & Kubany, E. S. (1995). Content validity in psychological assessment. *Psychological Assessment*, *7*(3), 238–247. https://doi.org/ct857f

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185. https://doi.org/d4fgnq

Hosmer, D. W., Jr., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). John Wiley & Sons.

Jolliffe, D., Farrington, D. P., Piquero, A. R., et al. (2017). Systematic review of self-reported delinquency and official offending. *Legal and Criminological Psychology*, *22*(1), 1-17.

Jones, A. P., Happé, F. G., Gilbert, F., et al. (2010). Feeling, caring, knowing: Boys with psychopathic tendencies and autism spectrum disorder. *Journal of Child Psychology and Psychiatry*, *51*(11), 1188-1197.

Kaiser, H. F. (1970). A second generation little jiffy. *Psychometrika*, *35*(4), 401–415. https://doi.org/bm98j6

Kaufman, J., Birmaher, B., Brent, D., Rao, U., Flynn, C., Moreci, P., Williamson, D., & Ryan, N. (1997). Schedule for Affective Disorders and Schizophrenia for School-Age Children-Present and Lifetime Version (K-SADS-PL). *Journal of the American Academy of Child & Adolescent Psychiatry*, *36*(7), 980–988. https://doi.org/c9h3n6

Kelley, T. L. (1927). *Interpretation of educational measurements*. Macmillan.

Kiddle, B., Inkster, B., Prabhu, G., Moutoussis, M., Whitaker, K. J., Bullmore, E. T., Dolan, R. J., Fonagy, P., Goodyer, I. M., & Jones, P. B. (2018). Cohort profile: The NeuroScience in Psychiatry Network (NSPN) 2400 cohort. *International Journal of Epidemiology*, *47*(1), 18–19. https://doi.org/gcs8rv

Kimonis, E. R., Frick, P. J., Skeem, J. L., Marsee, M. A., Cruise, K., Munoz, L. C., Aucoin, K. J., & Morris, A. S. (2008). Assessing callous–unemotional traits in adolescent offenders. *International Journal of Law and Psychiatry*, *31*(3), 241–252. https://doi.org/bjv898

Krueger, R. F., Markon, K. E., Patrick, C. J., & Iacono, W. G. (2005). Externalizing psychopathology in adulthood. *Journal of Abnormal Psychology*, *114*(4), 537–550. https://doi.org/bhj786

Lahey, B. B., Applegate, B., Chronis, A. M., Jones, H. A., Williams, S. H., Loney, J., & Waldman, I. D. (2008). Child and Adolescent Dispositions Scale. *Journal of Abnormal Psychology*, *117*(1), 144–154. https://doi.org/bvbtqc

Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd ed.). John Wiley & Sons.

Loeber, R., Stouthamer-Loeber, M., Van Kammen, W. B., & Farrington, D. P. (1991). Initiation, escalation and desistance in juvenile offending. *The Journal of Criminal Law and Criminology*, *82*(1), 36-82.

Lubke, G. H., & Muthén, B. O. (2004). Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data. *Structural Equation Modeling*, *11*(4), 514–534. https://doi.org/fs4h4s

Luo, J., Zhang, B., & He, J. (2021). Psychometric properties of the Proposed Specifiers for Conduct Disorder (PSCD) scale in Chinese adolescents. *Frontiers in Psychology*, *12*, 660888. https://doi.org/gmfk9f

Marsh, H. W. (1994). Confirmatory factor analysis models of multidimensionality. *Applied Psychological Measurement*, *18*(1), 5–16. https://doi.org/fshjzt

Maydeu-Olivares, A. (2017). Assessing the size of model misfit of IRT models. *Psychometrika*, *82*(3), 607–628. https://doi.org/gbpqk2

McDonald, R. P. (1999). *Test theory: A unified treatment*. Psychology Press.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here: The reliability of unit-weighted scales. *Psychological Methods*, *23*(3), 412–433. https://doi.org/gdsg5m

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. https://doi.org/brz958

Moffitt, T. E. (1993). Adolescence-limited and life-course-persistent antisocial behavior. *Psychological Review*, *100*(4), 674–701. https://doi.org/d976n2

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*(4), 557–585. https://doi.org/bpv8j6

Neufeld, S. A. S., Dunn, V. J., Jones, P. B., Croudace, T. J., & Goodyer, I. M. (2017). Reduction in adolescent depression after contact with mental health services. *The Lancet Psychiatry*, *4*(2), 120–127. https://doi.org/f9shjv

Neufeld, S. A. S., Dunn, V. J., Jones, P. B., Stochl, J., & Goodyer, I. M. (2018). Measurement invariance of the SDQ and a new 8-item externalising subscale (ABQ). *Journal of Abnormal Child Psychology*, *46*(7), 1531–1541. https://doi.org/gc6zrv

Neufeld, S. A. S., St Clair, M., Brodbeck, J., Wilkinson, P. O., Goodyer, I. M., & Jones, P. B. (2024). Measurement invariance in longitudinal bifactor models: Review and application based on the p factor. *Assessment*, *31*(4), 774–793. https://doi.org/gtfpg3

Nevitt, J., & Hancock, G. R. (2001). Performance of bootstrapping approaches to model test statistics. *Structural Equation Modeling*, *8*(3), 353–377. https://doi.org/dvjtcc

Panchal, P., Chiu, K., & Neufeld, S. A. S. (2026). *The psychometric premium of multidimensional alignment: A systematic review and meta-analysis of adolescent antisocial behaviour scales* [Manuscript in preparation]. Department of Clinical Psychology, University of East Anglia.

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/bpv966

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting. *Developmental Review*, *41*, 71–90. https://doi.org/f83r2k

R Core Team. (2021). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2013). Multidimensionality and structural coefficient bias in structural equation modeling. *Educational and Psychological Measurement*, *73*(1), 5–26. https://doi.org/f4hktc

Revelle, W. (2020). *psych: Procedures for psychological, psychometric, and personality research* (Version 2.0.9) [Computer software]. https://CRAN.R-project.org/package=psych

Reynolds, C. R., & Kamphaus, R. W. (2015). *BASC-3: Behavior Assessment System for Children* (3rd ed.). Pearson.

Rhew, I. C., Simpson, K., Tracy, M., Vander Stoep, A., McCauley, E., & Kosterman, R. (2016). Criterion validity of the Patient Health Questionnaire-9 (PHQ-9). *Journal of Affective Disorders*, *193*, 156–161. https://doi.org/gc3hkt

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*(1), 77. https://doi.org/dqxpg4

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Applying bifactor statistical indices in the evaluation of psychological measures. *Psychological Methods*, *21*(2), 223–234. https://doi.org/f8nk6z

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(1), 1–36. https://doi.org/f38z6n

Sass, D. A., Schmitt, T. A., & Marsh, H. W. (2014). What happens when we vary the number of response options? *Educational and Psychological Measurement*, *74*(1), 147–171. https://doi.org/f5jbt3

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*(4), 507–514. https://doi.org/bf9nkv

Sawyer, S. M., Azzopardi, P. S., Wickremarathne, D., & Patton, G. C. (2018). The age of adolescence. *The Lancet Child & Adolescent Health*, *2*(3), 223–228. https://doi.org/gc6pg3

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177. https://doi.org/dxfhjt

Shi, D., DiStefano, C., McDaniel, H. L., & Jiang, Z. (2020). Examining chi-square test statistics and fit indices for latent variable models. *Structural Equation Modeling*, *27*(1), 54–73. https://doi.org/ghm552

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction. *Annual Review of Psychology*, *69*, 487-510.

St Clair, M. C., Neufeld, S., Jones, P. B., Fonagy, P., Bullmore, E. T., Dolan, R. J., et al. (2017). Characterising the latent structure and organisation of self-reported thoughts, feelings and behaviours in adolescents and young adults. PLoS ONE, 12(4), e0175381. https://doi.org/10.1371/journal.pone.0175381

Streiner, D. L., Norman, G. R., & Cairney, J. (2015). *Health measurement scales* (5th ed.). Oxford University Press.

Waschbusch, D. A., & Walters, A. S. (2006). Psychometric properties of the disruptive behavior disorder rating scale. *Journal of Autism and Developmental Disorders*, *36*(3), 351–360. https://doi.org/dpntvz

Wiedermann, W., & von Eye, A. (2022). *Statistics and data analysis for the behavioral sciences*. Routledge.

World Health Organization. (2023). *Adolescent health*. https://www.who.int/health-topics/adolescent-health

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, *3*(1), 32–35. https://doi.org/cpnt7j

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's , Revelle's , and McDonald's. *Psychometrika*, *70*(1), 123–133. https://doi.org/d4n9z3

Table 1. *Antisocial Behaviours Questionnaire (ABQ): Item Stems and Content Summaries*

Item	Stem	Content
ABQ1	I deliberately broke the rules or disobeyed people (e.g. parents, teachers or	Rule-breaking/disobedience
ABQ2	I stole things (e.g. from home or a shop or school).	Theft/stealing
ABQ3	I deliberately damaged property (e.g. broke windows or chairs or wrote graffiti or	Property damage/vandalism
ABQ4	I deliberately hurt or threatened someone (e.g. bullying or fighting).	Aggression/threats (fighting/bullying)
ABQ5	I skipped lessons/work, skived, or played truant from school /college/work	Truancy/skiving
ABQ6	I deliberately lied or cheated to get what I wanted.	Lying/cheating
ABQ7	I ran away from home (e.g. for half a day or overnight).	Ran away from home
ABQ8	I threatened or forced someone to give their money or other belongings to me.	Extortion/robbery (threaten/force to give
ABQ9	I broke into someone else's property (e.g. into a house or a car).	Breaking & entering (burglary)
ABQ10	I have carried or used a weapon in a fight (e.g. a knife or a stick).	Weapon carrying/used in fight
ABQ11	I have deliberately hurt or been cruel to an animal (e.g. a pet).	Animal cruelty

Note. Stems are reproduced (abridged) from the original 11-item Behaviour Checklist used in ROOTS/NSPN; the "Content" labels are provided for cross-reference in subsequent analyses. For the present analyses, items were scored dichotomously (0 = no endorsement, 1 = any endorsement) and summed to yield a total score (0–8). Endorsement percentages are reported separately in Table 4.

Table 2.
Confirmatory Factor Analysis
Panel A. Confirmatory Factor Analysis Fit Indices for the One-Factor ABQ Model

Dataset	Estimator	χ^2 (df)	*p*	CFI	TLI	RMSEA [90% CI]	SRMR	*n*
ROOTS split half2	WLSMV (main) [1]	22.09 (20)	.34	.999	.998	.013 [.000, .039]	.055	576
NSPN 14–15	WLSMV	15.10 (20)	.77	1.000	1.020	.000 [.000, .000]	.103	486

Note. All models specified a single latent factor representing general antisocial behaviour using eight binary ABQ items. Goodness-of-fit values reflect robust estimators appropriate for categorical indicators. [1] DWLS was used only for bootstrap estimation of confidence intervals reported in Table S5.

Panel B. Standardised Factor Loadings for the One-Factor ABQ Model

Item	ROOTS split half2 (n = 576)	NSPN 14–15 (n = 486)
ABQ1 (Rule-breaking/disobedience)	.87	.78
ABQ2 (Theft/stealing)	.87	.71
ABQ3 (Property damage/vandalism)	.94	.85
ABQ4 (Aggression/threats [fighting/bullying])	.83	.72
ABQ5 (Truancy/skiving)	.80	.67
ABQ6 (Lying/cheating)	.80	.66
ABQ7 (Ran away from home)	.79	.48
ABQ10 (Weapon carrying/used in fight)	.70	.60

Note. Standardised loadings are based on a one-factor confirmatory factor analysis estimated using WLSMV with categorical indicators. All loadings were statistically significant ($p < .001$). Values for ROOTS split half2 are from the primary model; NSPN loadings reflect replication in the 14–15-year-old subsample.

Table 3.

Measurement Invariance Across Gender

Panel A. Gender Measurement Invariance Within Consolidated Developmental Age Bands (NSPN Baseline)

Age band	N (Fem/Male)	Invariance Step	χ^2 (df), p	CFI	TLI	RMSEA [90% CI]	SRMR	ΔCFI	ΔRMSEA	Decision
14–17 years	526 / 448	Configural	58.67 (40), .028	.986	.980	.031 [.010, .047]	.085	—	—	Accept
		Scalar	66.36 (46), .026	.984	.981	.030 [.011, .045]	.091	−.001	−.001	Accept
		Strict	98.95 (54), <.001	.965	.964	.041 [.028, .054]	.101	−.019	+.011	Accept†
18–24 years	752 / 652	Configural	71.09 (40), .002	.986	.980	.033 [.020, .046]	.080	—	—	Accept
		Scalar	110.11 (46), <.001	.970	.964	.045 [.034, .055]	.093	−.015	+.012	Accept*
		Strict	138.98 (54), <.001	.961	.959	.047 [.038, .057]	.111	−.009	+.002	Accept

Panel B. Latent Mean Differences (Reference = Female)

Age Band	ΔM	SE	z	p
14–17 years	−0.469	0.193	−2.43	.015
18–24 years	+0.028	0.136	0.21	.836

Note. Panel A presents gender measurement invariance within consolidated developmental age bands at NSPN baseline using multi-group CFA with WLSMV estimation and binary indicators. Consistent with recommendations for categorical data, metric invariance was bypassed and testing proceeded directly from configural to scalar invariance (Neufeld et al., 2024). Invariance decisions were based primarily on ΔCFI (< .010) and ΔRMSEA (< .007), with χ^2 statistics reported for completeness. Panel B shows latent mean differences with females as the reference group (latent mean = 0). †Strict invariance was retained where absolute model fit remained acceptable despite marginal Δ indices.

Table 4.

ROC operating characteristics at selected thresholds (ROOTS Baseline)

Criterion (K-SADS-PL)	Wave	AUC	95% CI	Youden cut-off	Sensitivity	Specificity
Conduct Disorder	Baseline (concurrent)	0.854	[0.823, 0.886]	≥ 2	0.846	0.739
Externalising Disorder	Baseline (concurrent)	0.768	[0.731, 0.804]	≥ 2	0.714	0.725

Note. ABQ totals use the 8-item binarised sum (0–8). The "≥ 2" cut-off corresponds to a pROC threshold of 1.5 (i.e., classify positive if ABQ ≥ 2). AUC 95% confidence intervals are based on the nonparametric DeLong method (DeLong et al., 1988). Operating points shown are those maximising Youden's index (J = Sensitivity + Specificity -1; Youden, 1950). Full threshold grids (1.0–8.0 in 0.5 steps), additional outcomes, and follow-up/predictive analyses are provided in Table S16.

Table 5.

Key Convergent, Divergent, and Predictive Validity of the ABQ Across Cohorts and Waves

Cohort	Wave	Measure	*r*	95% CI	*n*	Validity Role
ROOTS	baseline	APSD Total	.53	—	—	Convergent
		SDQ Hyperactivity	.32	—	—	Convergent (impulsivity)
		SDQ Prosocial	−.28	—	—	Divergent
NSPN	baseline	APSD Total	.53	[.50, .56]	2,375	Convergent
		APSD Narcissism	.38	[.34, .41]	2,349	Convergent
		APSD Impulsivity	.44	[.40, .47]	2,344	Convergent
		APSD Callousness	.26	[.22, .30]	2,350	Convergent
		ICU Total	.44	[.40, .47]	2,370	Convergent
		ICU Unemotional	.17	[.13, .21]	2,345	Convergent (attenuated)
		ICU Callous	.46	[.43, .50]	2,323	Convergent
		ICU Uncaring	.36	[.32, .39]	2,340	Convergent
		BIS Total	.41	[.38, .44]	2,356	Convergent (impulsivity)
		BIS Attention	.32	[.28, .36]	2,311	Convergent (impulsivity)
		BIS Motor	.26	[.22, .30]	2,300	Convergent (impulsivity)
		BIS Self-Control	.31	[.28, .35]	2,299	Convergent (impulsivity)
		BIS Attentional Impulsivity	.36	[.32, .39]	2,254	Convergent (impulsivity)
		CADS Daring	.22	[.18, .26]	2,322	Convergent (sensation seeking)
		CADS Prosociality	−.28	[−.31, −.24]	2,298	Divergent
NSPN	W2	APSD Total	.53	[.50, .56]	1,677	Convergent
		ICU Total	.44	[.40, .47]	1,666	Convergent
		BIS Total	.38	[.34, .42]	1,664	Convergent
		CADS Daring	.20	[.16, .25]	1,628	Convergent (sensation seeking)
		CADS Prosociality	−.26	[−.30, −.21]	1,623	Divergent
NSPN	W3	APSD Total	.52	[.48, .57]	1,121	Convergent
		ICU Total	.42	[.37, .47]	1,115	Convergent
		BIS Total	.33	[.28, .38]	1,100	Convergent (impulsivity)

Cohort	Wave	Measure	*r*	95% CI	*n*	Validity Role
		CADS Daring	.17	[.11, .22]	1,098	Convergent (sensation seeking)
		CADS Prosociality	−.30	[−.35, −.24]	1,083	Divergent
ROOTS	baseline → W3	DASI Total	.151	[.094, .207]	1155	Predictive

Note. Correlations are Pearson's r (95% CIs); DASI rows report Spearman's ρ. n denotes pairwise-complete cases. Negative Prosocial coefficients indicate inverse associations with antisocial behaviour. DASI analyses are prospective (baseline ABQ → ROOTS Wave 3). Measures: SDQ; APSD; ICU; BIS-11; CADS; DASI (see References). Supplementary Materials

Table S1.

Sample Characteristics and Retention by Cohort and Wave

Cohort	Wave	Wave Timing	N (analytic)	Retention (%)	Age M (Range)	% Female	Measures Available
ROOTS	W1	Baseline	1,153	100%	15.5 (14–16)	55.4	ABQ, SDQ, APSD, K-SADS
	W2	~18 months	874	75.8%	17.5 (16–18)	57.7	ABQ, SDQ
	W3	~36 months	999*	86.6%	19.5 (18–20)	57.4	ABQ, K-SADS, DASI
NSPN	W1	Baseline	2,378	100%	19.5 (14–25)	53.7	ABQ, APSD, ICU, BIS
	W2	~12 months	1,680	70.6%	20.5 (15–26)	57.3	ABQ, APSD, ICU, BIS
	W3	~24 months	1,127	47.4%	21.5 (16–27)	60.9	ABQ, APSD, ICU, BIS

Note. N (analytic) reflects the number of participants with valid ABQ data at each wave. ROOTS baseline was randomly split into df1 (n = 577) and df2 (n = 576) for exploratory and confirmatory analyses, with the full sample used for measurement invariance testing. NSPN utilized a stratified age-band sampling design (14–24y at baseline). For gender-by-age measurement invariance testing in NSPN, age bins were consolidated into mid-adolescent (14–17y) and young adult (18–24y) strata to maintain $N > 200$ per cell across longitudinal follow-ups.

Abbreviations: ABQ = Antisocial Behaviours Questionnaire; APSD = Antisocial Process Screening Device; BIS-11 = Barratt Impulsiveness Scale—11; CADS = Child and Adolescent Dispositions Scale; DASI = Drug, Alcohol and Self-Injury Scale; ICU = Inventory of Callous–Unemotional Traits; K-SADS-PL = Kiddie Schedule for Affective Disorders and Schizophrenia—Present and Lifetime; SDQ (parent) = Strengths and Difficulties Questionnaire, parent report.

Table S2.

Overview of Samples and Analyses

Analysis type	Cohort(s) / Wave(s) used	Description / purpose
Exploratory Factor Analysis (EFA)	ROOTS baseline (df1 n = 577)	Identify latent factor structure
Confirmatory Factor Analysis (CFA)	ROOTS baseline (df2 n = 576 NSPN baseline (14–15 y subsample)	Cross-validation and external replication
Measurement Invariance (MI)	ROOTS baseline–W3; NSPN baseline–W3	Test stability across gender, age, and time
Reliability (α, ω)	All waves, both cohorts	Assess internal consistency
Construct & criterion validity	ROOTS baseline, W3; NSPN baseline–W3	Convergent/divergent validity and diagnostic accuracy (ROC)

Note. ROOTS baseline was randomly split into df1 and df2 for EFA and CFA, respectively. *N*s shown in Table S1.

Table S3. *Comparator measures used in validity analyses: subscales, response formats, waves, and role*

Measure	Items	Study / Waves	Subscales (items)	Response scale	Role in current study
Strengths and Difficulties Questionnaire (SDQ; parent)	25	ROOTS: Baseline, ~18 mo	Prosocial (5), Hyperactivity (5), Emotional (5), Conduct (5), Peer (5)	0–2 (Not true–Certainly true)	Prosocial = divergent; Conduct = convergent
Antisocial Process Screening Device (APSD)	17	NSPN: Baseline, ~12 mo, ~24 mo	Callous–Unemotional (5), Impulsivity (5), Narcissism (7)	0–2 (Not at all–Definitely true)	Convergent
Child and Adolescent Dispositions Scale (CADS)	27	NSPN: Baseline, ~12 mo, ~24 mo	Prosociality (13), Negative Emotionality (9), Daring (5)	1–4 (Not at all–Very much)	Prosociality = divergent NE & Daring = convergent
Inventory of Callous–Unemotional Traits (ICU)	24	NSPN: Baseline, ~12 mo, ~24 mo	Callousness (11), Uncaring (8), Unemotional (5)	0–3 (Not at all–Definitely true)	Convergent
Barratt Impulsiveness Scale—11 (BIS-11)	25	NSPN: Baseline, ~12 mo, ~24 mo	Attentional, Motor, Non-planning	1–4 (Rarely/Never–Almost always)	Convergent
Drug, Alcohol and Self-Injury Scale (DASI)	10	ROOTS: Baseline, ~36 mo	Drug use, Alcohol use, Self-injury	0–3 (Never–Every day/nearly every day)	Predictive validity
K-SADS-PL (semi-structured interview)[a]	—	ROOTS: Baseline, ~36 mo	Diagnostic modules	Interview	Diagnostic/criterion (CD; externalising composites)

Note. This table summarises comparator measures from ROOTS and NSPN used to evaluate ABQ external validity. For each measure, we report total items, administration mode, subscales, response format, and role in the current study. Except for K-SADS-PL (diagnostic interview), all measures are self-reports. [a]K-SADS-PL is a semi-structured diagnostic interview comprising a screen interview and diagnostic supplements and can generate ~35 child and adolescent psychiatric diagnoses. Sources: SDQ (Goodman, 2001); APSD (Frick & Hare, 2001); CADS (Lahey et al., 2008); ICU (Essau et al., 2006; Kimonis et al., 2008); BIS-11 (Patton et al., 1995); DASI (Kiddle et al., 2018); K-SADS-PL (Kaufman et al., 1997).

Table S4.

ABQ Item Content and Baseline Endorsement (%) in ROOTS and NSPN Cohorts

Item	Content	Baseline Endorsement (%)			
		ROOTS			NSPN
		Full	df1 (EFA)	df2 (CFA)	14-15yo (EFA)
ABQ1	Rule-breaking/disobedience	42.1	43.0	41.1	38.1
ABQ2	Theft/stealing	8.8	10.4	7.1	7.7
ABQ3	Property damage/vandalism	6.2	6.9	5.6	2.6
ABQ4	Aggression/threats (fighting/bullying)	12.5	12.5	12.5	8.5
ABQ5	Truancy/skiving	6.3	5.9	6.8	8.7
ABQ6	Lying/cheating	24.1	26.0	22.2	28.9
ABQ7	Ran away from home	3.5	4.0	3.0	2.2
ABQ8	Extortion/robbery (threaten/force to give belongings)	*0.4	—	—	—
ABQ9	Breaking & entering (burglary)	*0.6	—	—	—
ABQ10	Weapon carrying/used in fight	2.6	2.8	2.4	0.6
ABQ11	Animal cruelty	*1.6	—	—	—

Note. Percentages reflect the proportion endorsing any behaviour after binarisation (0 = never; 1 = sometimes/mostly/always). ROOTS baseline totals: Full N = 1,153; df1 (EFA) n = 577; df2 (CFA) n = 576. NSPN baseline (14–15 y): N = 486; endorsement percentages are presented assuming missing responses were removed so that a common denominator of 486 applies to all items. Asterisks (*) mark items with < 2% endorsement at ROOTS baseline; these ultra–low-prevalence items (ABQ8, ABQ9, ABQ11) were excluded from psychometric modelling. Em dashes (—) indicate items excluded from subsequent analyses, so no percentages are reported.

Table S5.

Tetrachoric correlations among ABQ items (ROOTS df1, n = 577; final 8-item set)

	ABQ1	ABQ2	ABQ3	ABQ4	ABQ5	ABQ6	ABQ7	ABQ10
ABQ1 (Rule-breaking/disobedience)	—							
ABQ2 (Theft/stealing)	.55	—						
ABQ3 (Property damage/vandalism)	.72	.63	—					
ABQ4 (Aggression/threats [fighting/bullying])	.64	.51	.68	—				
ABQ5 (Truancy/skiving)	.46	.57	.55	.51	—			
ABQ6 (Lying/cheating)	.68	.48	.60	.51	.49	—		
ABQ7 (Ran away from home)	.47	.56	.47	.56	.71	.46	—	
ABQ10 (Weapon carrying/used in fight)	.62	.61	.59	.67	.56	.34	.28	—

Note. Values are tetrachoric correlations (lower triangle) among the eight retained ABQ items. KMO = .69**;** Bartlett's $\chi^2(28) = 3205.60$, $p < .001$**.**

Table S6.

Tetrachoric correlations among ABQ items (ROOTS df2, n = 576; final 8-item set)

	ABQ1	ABQ2	ABQ3	ABQ4	ABQ5	ABQ6	ABQ7	ABQ10
ABQ1 (Rule-breaking/disobedience)	—							
ABQ2 (Theft/stealing)	.80	—						
ABQ3 (Property damage/vandalism)	.80	.76	—					
ABQ4 (Aggression/threats [fighting/bullying])	.70	.64	.84	—				
ABQ5 (Truancy/skiving)	.68	.61	.77	.66	—			
ABQ6 (Lying/cheating)	.69	.78	.68	.59	.64	—		
ABQ7 (Ran away from home)	.64	.72	.62	.69	.73	.60	—	
ABQ10 (Weapon carrying/used in fight)	.61	.56	.68	.62	.50	.60	.38	—

Note. Values are tetrachoric correlations (lower triangle) among the eight retained ABQ items. KMO = .80**;** Bartlett's $\chi^2(28) = 4353.66$, $p < .001$**.**

Table S7.

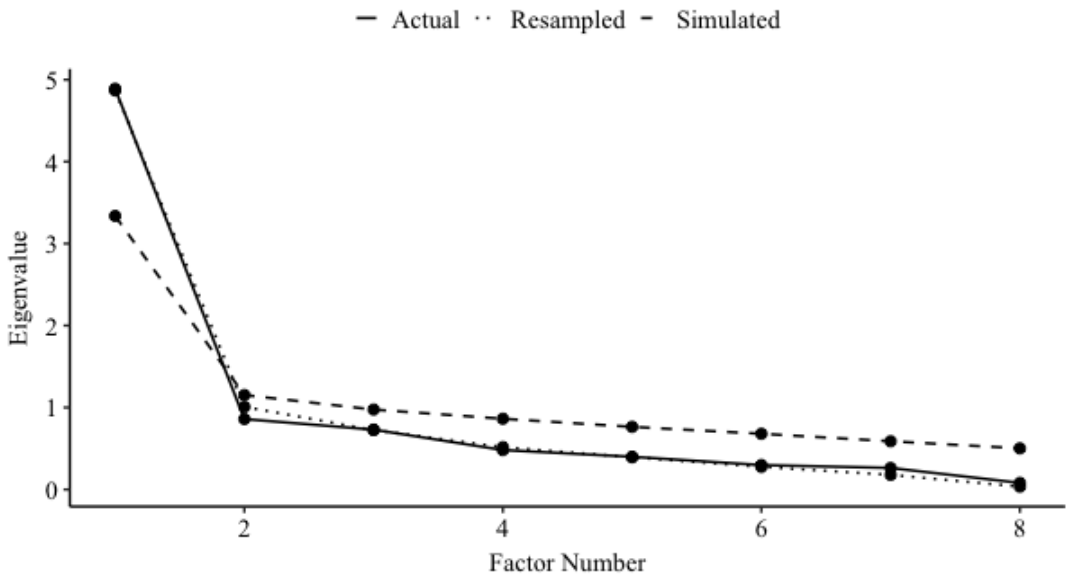
Tetrachoric correlations among ABQ items (NSPN 14–15 years, n ≈ 487; final 8-item set)

	ABQ1	ABQ2	ABQ3	ABQ4	ABQ5	ABQ6	ABQ7	ABQ10
ABQ1 (Rule-breaking/disobedience)	—							
ABQ2 (Theft/stealing)	.44	—						
ABQ3 (Property damage/vandalism)	.28	.23	—					
ABQ4 (Aggression/threats [fighting/bullying])	.49	.24	.34	—				
ABQ5 (Truancy/skiving)	.53	.46	.21	.27	—			
ABQ6 (Lying/cheating)	.64	.56	.29	.54	.51	—		
ABQ7 (Ran away from home)	.30	.09	.34	.40	.52	.51	—	
ABQ10 (Weapon carrying/used in fight)	.26	.37	.34	.62	.35	.35	.37	—

Note. Values are tetrachoric correlations (lower triangle) among the eight retained ABQ items (ABQ1–ABQ7 and ABQ10 Weapon) in the NSPN 14–15-year-old subsample. All coefficients are positive and generally moderate in magnitude. KMO = .69**;** Bartlett's $\chi^2(28) = 1564.06$, $p < .001$**.**

Figure S1.

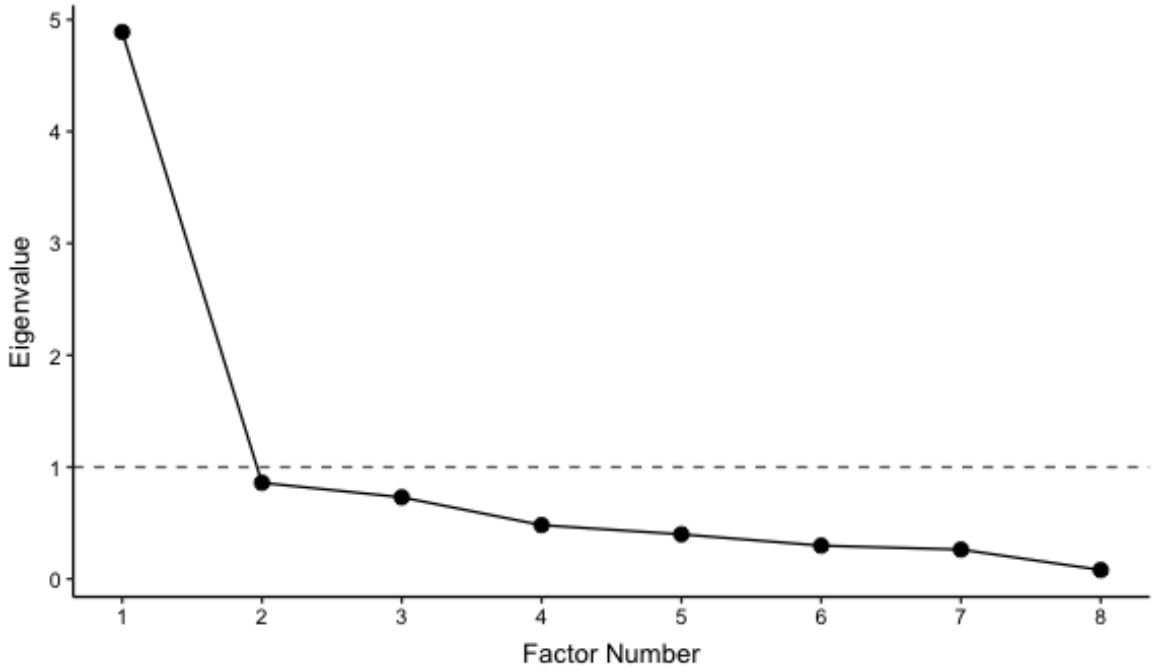
Parallel Analysis Plot for the ABQ (ROOTS splithalf1).



Note. The plot shows eigenvalues from the observed data (solid line), simulated random data (dashed line), and bootstrap resamples (dotted line). Factors are retained where the observed eigenvalues exceed both comparison lines. Although parallel analysis suggested up to three factors, convergent evidence from the Scree test and Velicer's MAP supported a one-factor solution.

Figure S2.

Scree Plot of Eigenvalues for ABQ (ROOTS splithalf1)



Note. The scree plot shows a steep drop after the first eigenvalue (≈ 5.0), followed by a clear elbow, indicating a dominant single factor. Subsequent eigenvalues fall below 1.0 (dashed line, Kaiser criterion), supporting retention of one factor. The analysis was based on tetrachoric correlations among the eight retained ABQ items.

Table S8.

Exploratory Factor Analysis of the ABQ (ROOTS df1, n = 577; final 8-item set)

Item	Standardised Loading (MR1)	Communality (h^2)	Uniqueness (u^2)	Complexity
ABQ1 (Rule-breaking/disobedience)	.81	.65	.35	1
ABQ2 (Theft/stealing)	.75	.56	.44	1
ABQ3 (Property damage/vandalism)	.83	.69	.31	1
ABQ4 (Aggression/threats [fighting/bullying])	.79	.62	.38	1
ABQ5 (Truancy/skiving)	.73	.53	.47	1
ABQ6 (Lying/cheating)	.68	.46	.54	1
ABQ7 (Ran away from home)	.66	.44	.56	1
ABQ10 (Weapon carrying/used in fight)	.71	.50	.50	1

Note. Extraction method = minimum residual (minres); rotation = oblimin. Model fit: RMSR = .08, off-diagonal fit = .98. SS loadings = 4.46, explaining 56% of common variance. Factor-score adequacy: r = .96, R^2 = .92, minimum possible correlation = .84. Based on tetrachoric correlation matrix of eight binary ABQ items (ROOTS baseline, df1; n = 577).

Table S9.

Bootstrapped 95% Confidence Intervals for Standardised Residual Variances for the One-Factor ABQ Model

Item	ROOTS splithalf2 (*n* = 576)	95% CI	NSPN 14–15 (*n* = 486)	95% CI
ABQ1 (Rule-breaking/disobedience)	0.262	[0.145, 0.370]	0.498	[0.273, 0.801]
ABQ2 (Theft/stealing)	0.245	[0.120, 0.384]	0.636	[0.243, 0.843]
ABQ3 (Property damage/vandalism)	0.119	[−0.008, 0.255]	0.699	[0.331, 0.902]
ABQ4 (Aggression/threats [fighting/bullying])	0.297	[0.173, 0.431]	0.514	[0.230, 0.760]
ABQ5 (Truancy/skiving)	0.347	[0.199, 0.531]	0.555	[0.300, 0.810]
ABQ6 (Lying/cheating)	0.352	[0.225, 0.480]	0.350	[0.115, 0.665]
ABQ7 (Ran away from home)	0.366	[0.140, 0.639]	0.571	[0.182, 0.837]
ABQ10 (Weapon carrying/used in fight)	0.508	[0.270, 0.755]	0.593	[0.183, 0.872]

Note. Values are based on 1,000 bootstrap samples using DWLS estimation. CIs are percentile confidence intervals derived from the bootstrapped sampling distribution of standardised residual variances (Std.All) in one-factor ABQ CFA models for ROOTS df2 and NSPN 14–15 datasets.

Table S10.

Gender Measurement Invariance of the ABQ Across ROOTS and NSPN

Panel A. Wave-Specific Gender Invariance Models (Snapshot Analyses)

Cohort	Wave (Age)	Invariance Step	χ^2 (df), p	CFI	TLI	RMSEA [90% CI]	SRMR	ΔCFI	ΔRMSEA	ΔM	Decision
ROOTS	Baseline (~14.5y)	Configural	49.59 (40), .142	.996	.995	.020 [.000, .037]	.066	—	—	—	Accept
		Scalar	70.19 (46), .012	.991	.988	.030 [.014, .044]	.077	−.005	+.010	−0.205	Accept
		Strict	91.46 (54), .001	.985	.985	.035 [.022, .047]	.079	−.005	+.004	−0.205	Accept
	W2 (~16y)	Configural	27.90 (40), .926	1.000	1.012	.000 [.000, .011]	.060	—	—	—	Accept
		Scalar	48.17 (46), .385	.998	.998	.010 [.000, .034]	.070	−.002	+.010	−0.051	Accept
		Strict	48.17 (54), .385	.998	.998	.010 [.000, .034]	.070	.000	.000	−0.051	Accept
	W3 (~17.5y)	Configural	206.55 (40), <.001	.903	.865	.091 [.079, .104]	.136	—	—	—	Accept†
		Scalar	219.39 (46), <.001	.899	.877	.087 [.076, .099]	.136	−.004	−.004	+0.173	Accept†
		Strict	219.39 (54), <.001	.899	.877	.087 [.076, .099]	.136	.000	.000	+0.173	Accept†
NSPN	W1 (14–24y)	Configural	66.90 (40), .005	.991	.987	.024 [.013, .034]	.061	—	—	—	Accept
		Scalar	89.46 (46), <.001	.985	.982	.028 [.019, .037]	.070	−.006	+.004	−0.326*	Accept
		Strict	96.72 (54), <.001	.985	.985	.026 [.017, .034]	.073	.000	−.002	−0.321*	Accept
	W2 (15–25y)	Configural	45.19 (40), .264	.998	.997	.012 [.000, .028]	.062	—	—	—	Accept
		Scalar	67.91 (46), .019	.990	.987	.024 [.010, .035]	.067	−.008	+.012	−0.405*	Accept
		Strict	78.27 (54), .017	.988	.988	.023 [.010, .034]	.069	−.002	−.001	−0.390*	Accept
	W3 (16–26y)	Configural	40.09 (40), .466	1.000	1.000	.002 [.000, .029]	.062	—	—	—	Accept
		Scalar	50.74 (46), .292	.998	.998	.014 [.000, .032]	.068	−.002	+.012	−1.303*†	Accept
		Strict	74.39 (54), .034	.993	.993	.026 [.007, .039]	.089	−.005	+.012	−0.369*	Accept

Panel B. Longitudinal Gender Measurement Invariance Across Waves 1–3 Using Single Wide-by-Time Multi-Group Models

Cohort	Invariance Step	N	χ^2 (df), p	CFI	TLI	RMSEA [90% CI]	SRMR	ΔCFI	ΔRMSEA	Decision
ROOTS	Configural	1,238	796.83 (498), <.001	.919	.910	.039 [.034, .044]	.134	—	—	Accept
	Scalar		803.95 (516), <.001	.922	.916	.038 [.033, .043]	.135	+.003	−.001	Accept
	Strict		805.50 (540), <.001	.928	.926	.036 [.030, .041]	.139	+.006	−.002	Accept
NSPN	Configural	2,403	676.82 (516), <.001	.974	.972	.016 [.013, .019]	.131	—	—	Accept
	Scalar		676.82 (516), <.001	.974	.972	.016 [.013, .019]	.131	.000	.000	Accept
	Strict		704.07 (540), <.001	.973	.972	.016 [.012, .019]	.137	−.001	.000	Accept

Note. WLSMV estimation with binary indicators; the metric stage was bypassed and invariance tested directly from configural to scalar to strict (Neufeld et al., 2024). Panel A presents wave-specific ("snapshot") gender invariance models estimated separately at each wave. Panel B presents single longitudinal wide-by-time multi-group models testing gender invariance simultaneously across Waves 1–3 within each cohort, accounting for missingness over time. Measurement invariance decisions were based primarily on ΔCFI < .010 and ΔRMSEA < .007; χ^2 statistics are reported for completeness but were not primary decision criteria.

Table S11.

Measurement Invariance Across Age bands

Panel A. Age-Band Measurement Invariance of the ABQ at NSPN Baseline

Invariance Step	χ^2 (df), p	CFI	TLI	RMSEA [90% CI]	SRMR	ΔCFI	ΔRMSEA	Decision
Configural	100.57 (80), .060	.994	.992	.023 [.000, .036]	.080	—	—	Accept
Scalar	128.96 (98), .020	.991	.990	.026 [.011, .037]	.090	−.003	+.003	Accept
Strict	164.66 (122), .006	.988	.989	.027 [.015, .037]	.113	−.003	+.001	Accept

Panel B. Latent Mean Differences (reference = 14-15 years)

Wave	Age Band	Mean Difference (ΔM)	p
Baseline	14–15 years	0.00	—
	16–17 years	+0.366	.036
	18–19 years	–0.605	.011
	20–21 years	–0.678	.006
	22–24 years	–0.079	.614

Note. Panel A presents age-band measurement invariance of the ABQ at NSPN baseline after excluding the 14–15 age group to evaluate stability of invariance estimates, as low behavioural endorsement in early adolescence produced parameter instability in the full model. Multi-group CFA used WLSMV estimation with binary indicators; metric invariance was bypassed and testing proceeded directly from configural to scalar invariance (Neufeld et al., 2024). Invariance decisions were based primarily on ΔCFI (< .010) and ΔRMSEA (< .007). Panel B reports latent mean differences from the baseline scalar model estimated across the full developmental range (14–24 years), with the 14–15 age band serving as the reference group (latent mean = 0).

Table S12.

Reliability and Factor Adequacy of the 8-item ABQ (N = 1,153)

Category	Statistic	Estimate [95% CI]
Internal Consistency	Cronbach's α	.73 [.71, .76]
	McDonald's ω_{Total}	.93
	Guttman's λ_6	.74
	Mean Inter-item Correlation	.28
Factor Adequacy	Explained Common Variance (ECV)	1.00
	Correlation of Scores with Factor	.97
	Multiple R^2 of Scores with Factor	.93
	Min. Correlation of Factor Scores	.87

Note. Estimates derived from tetrachoric correlations (ROOTS baseline). Values indicate strong internal consistency for a single-factor, eight-item model. In this single-factor model, $\omega_{Hierarchical}$ is effectively equivalent to ω_{Total}. ECV of 1.00 confirms absolute unidimensionality.

Table S13.

Item-Level Reliability Diagnostics for the ABQ (ROOTS N = 1,153)

Item	Mean	SD	$r_{(total)}$	$r_{(drop)}$	$\alpha_{(if\ deleted)}$
ABQ1 (Rule-breaking/disobedience)	.42	.49	.72	.48	.71
ABQ2 (Theft/stealing)	.09	.28	.62	.48	.70
ABQ3 (Property damage/vandalism)	.06	.24	.63	.51	.69
ABQ4 (Aggression/threats [fighting/bullying])	.12	.33	.67	.51	.69
ABQ5 (Truancy/skiving)	.06	.24	.56	.43	.71
ABQ6 (Lying/cheating)	.24	.43	.70	.50	.69
ABQ7 (Ran away from home)	.04	.18	.46	.36	.72
ABQ10 (Weapon carrying/used in fight)	.03	.16	.40	.31	.73

Note. Tetrachoric correlations used. *r*(total) and *r*(drop) represent item–total and corrected item–total correlations. No item removal increases the overall α beyond .73.

Table S14.

Full ROC Threshold Grid for ABQ Predicting K-SADS Conduct Disorder (Baseline ROOTS)

ABQ threshold	Sensitivity	Specificity	Youden's J	AUC (95% CI)
1.0	.923	.527	.450	.854 [.823, .886]
2.0	.846	.739	.585	.854 [.823, .886]
3.0	.692	.869	.561	.854 [.823, .886]
4.0	.538	.932	.470	.854 [.823, .886]
5.0	.308	.963	.271	.854 [.823, .886]
6.0	.231	.981	.212	.854 [.823, .886]
7.0	.154	.991	.145	.854 [.823, .886]
8.0	.000	1.000	.000	.854 [.823, .886]

Note. ABQ = Antisocial Behaviours Questionnaire (0–8 sum, binarised items). K-SADS = Kiddie Schedule for Affective Disorders and Schizophrenia (Present and Lifetime Version). AUC 95% CIs from the non-parametric DeLong method (DeLong et al., 1988); optimal operating points may be selected via Youden's J = sensitivity + specificity – 1 (Youden, 1950). Values shown here are concurrent ROC for ROOTS baseline only.

Table S15.

Full ABQ Convergent, Divergent, and Predictive Validity Across Cohorts and Waves

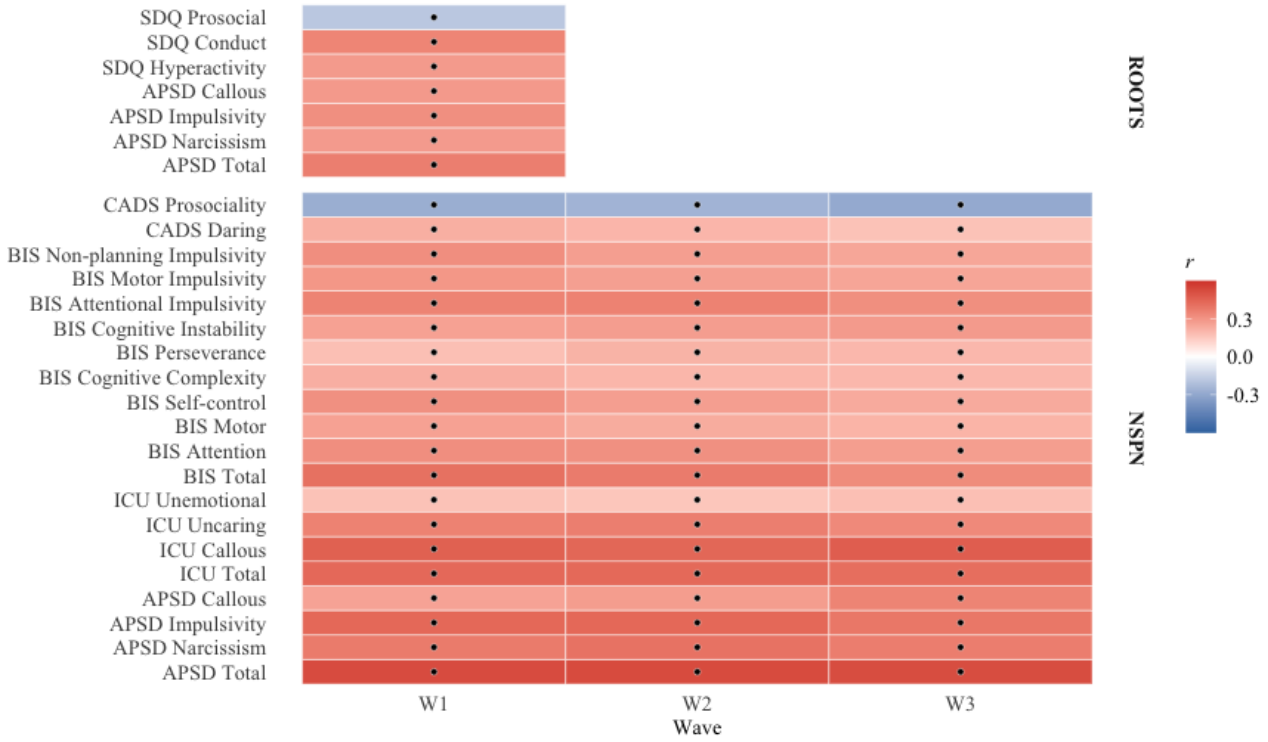
Cohort	Wave	Measure	*r*	95% CI	*n*
ROOTS	baseline	APSD Total	.53	—	—
		SDQ Hyperactivity	.32	—	—
		SDQ Prosocial	−.28	—	—
NSPN	baseline	APSD Total	.528	[.499, .557]	2,375
		APSD Narcissism	.378	[.343, .412]	2,349
		APSD Impulsivity	.438	[.404, .470]	2,344
		APSD Callous	.258	[.220, .295]	2,350
		ICU Total	.437	[.404, .469]	2,370
		ICU Unemotional	.167	[.127, .206]	2,345
		ICU Callous	.463	[.431, .495]	2,323
		ICU Uncaring	.357	[.322, .392]	2,340
		BIS Total	.409	[.375, .442]	2,356
		BIS Attention	.320	[.283, .356]	2,311
		BIS Motor	.263	[.224, .300]	2,300
		BIS Self-control	.313	[.276, .350]	2,299
		BIS Cognitive Complexity	.225	[.186, .264]	2,320
		BIS Perseverance	.177	[.136, .216]	2,269
		BIS Cognitive Instability	.263	[.224, .301]	2,288
		BIS Attentional Impulsivity	.355	[.319, .391]	2,254
		BIS Motor Impulsivity	.295	[.256, .332]	2,234
		BIS Non-planning Impulsivity	.320	[.283, .356]	2,265
		CADS Prosociality	−.276	[−.314, −.238]	2,298
		CADS Daring	.223	[.184, .262]	2,322
NSPN	W2	APSD Total	.530	[.495, .564]	1,677
		APSD Narcissism	.404	[.363, .443]	1,661
		APSD Impulsivity	.439	[.399, .477]	1,657
		APSD Callous	.277	[.232, .321]	1,666
		ICU Total	.435	[.395, .473]	1,666
		ICU Unemotional	.158	[.111, .205]	1,658
		ICU Callous	.443	[.403, .482]	1,621
		ICU Uncaring	.368	[.325, .409]	1,639
		BIS Total	.377	[.335, .418]	1,664
		BIS Attention	.315	[.271, .358]	1,640
		BIS Motor	.232	[.186, .278]	1,626
		BIS Self-control	.272	[.226, .316]	1,639
		BIS Cognitive Complexity	.202	[.155, .248]	1,636
		BIS Perseverance	.212	[.165, .258]	1,619
		BIS Cognitive Instability	.276	[.231, .321]	1,631
		BIS Attentional Impulsivity	.358	[.315, .400]	1,608
		BIS Motor Impulsivity	.270	[.224, .315]	1,591
		BIS Non-planning Impulsivity	.273	[.227, .317]	1,613
		CADS Prosociality	−.257	[−.302, −.211]	1,623
		CADS Daring	.203	[.156, .249]	1,628
NSPN	W3	APSD Total	.524	[.480, .565]	1,121
		APSD Narcissism	.371	[.319, .420]	1,111
		APSD Impulsivity	.391	[.340, .439]	1,113
		APSD Callous	.351	[.298, .401]	1,108
		ICU Total	.419	[.370, .467]	1,115
		ICU Unemotional	.177	[.119, .233]	1,098
		ICU Callous	.470	[.422, .515]	1,089
		ICU Uncaring	.335	[.282, .386]	1,101
		BIS Total	.329	[.276, .381]	1,100
		BIS Attention	.274	[.218, .328]	1,091

Cohort	Wave	Measure	*r*	95% CI	*n*
		BIS Motor	.208	[.150, .264]	1,074
		BIS Self-control	.236	[.179, .292]	1,080
		BIS Cognitive Complexity	.198	[.140, .254]	1,084
		BIS Perseverance	.197	[.139, .253]	1,082
		BIS Cognitive Instability	.284	[.228, .338]	1,070
		BIS Attentional Impulsivity	.319	[.264, .372]	1,063
		BIS Motor Impulsivity	.254	[.197, .309]	1,057
		BIS Non-planning Impulsivity	.250	[.193, .305]	1,066
		CADS Prosociality	−.295	[−.348, −.240]	1,083
		CADS Daring	.167	[.109, .224]	1,098
ROOTS	W3	DASI Total	.151	[.094, .207]	1155
		DASI Drug Use	.220	[.164, .274]	1155
		DASI Alcohol Use	.085	[.027, .142]	1155
		DASI Self-Injury	.086	[.028, .143]	1155

Note. Pearson's r (95% CI) reported; DASI rows report Spearman's ρ. n reflects pairwise-complete cases. Negative coefficients for prosocial scales indicate inverse associations with antisocial behaviour. DASI analyses represent prospective predictive validity (Baseline ABQ predicting Wave 3 outcomes). Measures represent theoretically related constructs within the externalising domain; interpretation of convergent and divergent patterns is based on expected direction and magnitude of associations.

Figure S3.

ABQ-Comparator Correlations Across Cohorts and Waves



Note. Colour encodes the magnitude and sign of *r* (white = near zero; red/blue = stronger positive/negative). Dots mark coefficients whose 95% confidence intervals exclude zero. ROOTS: W1 = baseline; NSPN: Packs 1–3 mapped to W1–W3. Cells without estimates are shown in light grey.

Chapter 5. Discussion and Critical Analysis

This chapter integrates the findings from the systematic review (Chapter 2) and the empirical psychometric evaluation of the Antisocial Behaviour Questionnaire (Chapter 4). It begins by briefly summarising the principal findings of each paper before advancing a synthesis of their combined contribution to the measurement of adolescent antisocial behaviour. The chapter then critically evaluates the methodological strengths and limitations of the portfolio as a whole, including considerations relating to construct definition, psychometric standards, and developmental comparability. Finally, implications for research and applied contexts are discussed, and priorities for future investigation are outlined. Through this integrative analysis, the portfolio's broader contribution to advancing coherence and comparability in antisocial behaviour measurement is clarified.

Summary of Principal Findings

Systematic Review

The systematic review provided a structured evaluation of instruments designed to assess adolescent antisocial behaviour, integrating a 12-tier taxonomic audit with DSM-5 Conduct Disorder domain alignment (American Psychiatric Association, 2013) and COSMIN-based psychometric appraisal (Mokkink et al., 2018). Although numerous measures are available and widely used under varied labels (Krueger et al., 2005; Waschbusch & Walters, 2006), considerable heterogeneity was evident in construct definition, behavioural scope, and methodological quality. Sequential application of taxonomic and psychometric criteria demonstrated that only a small subset of instruments achieved both comprehensive coverage of the full behavioural range specified within DSM-5 Conduct Disorder (American Psychiatric Association, 2013) and moderate-to-high certainty under contemporary evaluation standards.

The review further indicated that validation practices within the antisocial behaviour literature reflect multiple methodological periods. Earlier studies frequently employed analytic approaches such as principal component analysis (Fabrigar et al., 1999) and relied primarily on Cronbach's alpha as an index of reliability, despite recognised limitations of this estimator (Dunn et al., 2014; Hayes & Coutts, 2020). More recent work has increasingly incorporated confirmatory factor analysis (Brown, 2015), alternative reliability estimators (Hayes & Coutts, 2020), and multi-group measurement invariance testing (Meredith, 1993; Putnick & Bornstein, 2016), reflecting broader developments in psychometric methodology. Variability in reporting and analytic depth therefore appears to represent the natural evolution of methodological expectations rather than systematic oversight.

Empirical Evaluation of the Antisocial Behaviour Questionnaire

The empirical study examined the psychometric properties of the 8-item Antisocial Behaviour Questionnaire (ABQ) across two large UK cohorts spanning adolescence and emerging adulthood. Findings supported a robust unidimensional latent structure and strong internal consistency, with evidence of scalar and strict measurement invariance across gender and developmental stages (Meredith, 1993; Chen, 2007; Putnick & Bornstein, 2016) once endorsement sparsity was addressed. These results permit meaningful comparison of antisocial behaviour across key demographic and developmental groups (Meredith, 1993; Putnick & Bornstein, 2016).

Convergent and predictive associations with related externalising traits further supported construct validity within the broader externalising spectrum (Cronbach & Meehl, 1955; Krueger et al., 2005), and receiver operating characteristic analyses (Youden, 1950; Altman & Bland, 1994) indicated that the measure can function as a balanced screening index within community samples. Despite its brevity, the ABQ demonstrated coverage of behaviours spanning relatively common rule-breaking acts to less frequent, higher-severity behaviours,

positioning it as a concise yet behaviourally anchored dimensional measure. At the same time, sparse endorsement of certain high-threshold behaviours in categorical modelling contexts (Brown, 2015) highlighted the need for further validation in higher-risk or forensic populations.

Integrated Contribution of the Portfolio

Taken together, the two papers address complementary aspects of a shared problem within adolescent antisocial behaviour research: the proliferation of instruments without consistent criteria for evaluating construct coverage and psychometric adequacy (Marsh, 1994; DeVellis, 2016; Mokkink et al., 2018). The systematic review demonstrated that although numerous measures are available and widely used under varied nomenclature, relatively few simultaneously provide comprehensive behavioural representation and contemporary psychometric support. In doing so, the review established a structured framework for evaluating instruments through sequential taxonomic filtering, DSM-5 Conduct Disorder domain alignment (APA, 2013), and COSMIN-informed appraisal (Mokkink et al., 2018).

The empirical evaluation of the Antisocial Behaviour Questionnaire (ABQ) directly addressed the feasibility gap identified in the review. Specifically, it tested whether a brief DSM-aligned behavioural index could satisfy contemporary standards of structural validity, developmental invariance, and external validity within large community cohorts. By integrating latent modelling (Brown, 2015), multi-group invariance testing (Meredith, 1993; Putnick & Bornstein, 2016), and diagnostic threshold evaluation using ROC analyses (Youden, 1950; Altman & Bland, 1994), the study demonstrated that psychometric rigour and applied feasibility need not be mutually exclusive. The ABQ was therefore not merely assumed to be adequate on the basis of prior use or internal consistency but evaluated against the same sequential criteria established in the systematic review.

The portfolio therefore contributes more than the evaluation of individual instruments. It offers a reproducible approach to adolescent antisocial behaviour measurement that integrates conceptual clarity with psychometric rigour. The 12-tier taxonomic audit provides a means of delimiting the construct and distinguishing broadband antisocial behaviour from adjacent but conceptually distinct phenomena (Marsh, 1994; Haynes et al., 1995). DSM-5 Conduct Disorder domain mapping offers a pragmatic behavioural anchor to ensure comprehensive coverage across the developmental severity gradient (APA, 2013). Contemporary latent modelling and invariance testing provide safeguards against artefactual group or age differences (Meredith, 1993; Putnick & Bornstein, 2016).

Together, these components advance the field towards greater comparability, transparency, and alignment between instrument purpose and measurement standards. In this respect, the contribution of the portfolio lies not in proposing a single definitive measure, but in clarifying how adolescent antisocial behaviour instruments may be systematically evaluated and interpreted within both research and applied contexts.

Defining and Delimiting Antisocial Behaviour for Measurement

A central issue underlying both papers concerns how antisocial behaviour (AB) is defined and delimited for measurement purposes. Within developmental psychopathology and criminology, antisocial behaviour is typically conceptualised as a behavioural subset within the broader externalising spectrum (Krueger et al., 2005; Krueger & Markon, 2006; Lahey et al., 2012), defined as actions that violate social norms, legal rules, or the rights of others (American Psychiatric Association, 2013; Frick & Viding, 2009). This distinguishes it from adjacent constructs such as impulsivity, oppositionality, or general behavioural dysregulation, which may co-occur but do not necessarily involve norm violation or harm to others.

Conduct Disorder (CD) provides a categorical clinical formulation for minors (APA, 2013), defined by the presence of multiple antisocial symptoms causing impairment. However, AB extends beyond diagnostic thresholds and is better conceptualised as a dimensional continuum ranging from lower-frequency rule-breaking to more serious and legally consequential behaviours (Frick & Viding, 2009; Moffitt, 1993). Within this framework, CD represents a thresholded clinical endpoint, whereas AB captures the full behavioural distribution across adolescence and emerging adulthood.

Conceptual ambiguity often arises because measures labelled as assessing antisocial behaviour vary in breadth and construct level. Some focus narrowly on overt aggression; others integrate behavioural acts with dispositional traits such as callousness or unemotionality. Although psychopathic traits frequently co-occur with antisocial conduct and may increase risk for persistence (Frick et al., 2014; Hare & Neumann, 2008), they represent affective–interpersonal characteristics rather than behavioural violations per se. Conflation of these levels—acts versus traits—risks obscuring what is being measured and complicates cross-instrument comparability (Marsh, 1994; Haynes et al., 1995).

Developmental context further complicates construct definition. The age–crime curve, one of the most replicated findings in criminology, demonstrates that antisocial and criminal behaviours rise sharply in early adolescence, peak in the late teenage years, and decline during the transition to adulthood (Farrington, 2003; Moffitt, 1993). The initial increase corresponds with pubertal maturation, heightened peer influence, and a developmental imbalance between reward sensitivity and regulatory control (Steinberg, 2008; Shulman et al., 2016). The subsequent decline is associated with neurocognitive maturation and the acquisition of conventional adult roles that increase “stakes in conformity” (Sampson & Laub, 1993).

Moffitt's developmental taxonomy provides a theoretical account of this pattern by distinguishing adolescence-limited and life-course persistent pathways (Moffitt, 1993; Moffitt et al., 2002). From a measurement perspective, this distinction underscores the need to capture behavioural breadth without pathologizing normative adolescent experimentation.

This developmental heterogeneity introduces a methodological challenge. Apparent differences in developmental patterning may arise not only from population variation but also from the behavioural scope of the instrument. The 12-tier taxonomic audit applied within the systematic review was developed to bring greater transparency to these construct boundaries and reduce construct drift (Marsh, 1994; Haynes et al., 1995). Alignment with the four DSM-5 CD domains (APA, 2013) was subsequently used as a comparative anchor to ensure coverage across the full behavioural gradient.

Taken together, these considerations suggest that antisocial behaviour is best conceptualised as a developmentally dynamic, behaviourally defined construct situated within the broader externalising spectrum (Krueger et al., 2005). Clear delineation between behavioural acts and dispositional traits, attention to developmental patterning reflected in the age–crime curve, and transparency regarding behavioural coverage are therefore central to coherent assessment.

Psychometric Standards and Developmental Comparability

Clear construct definition alone is insufficient without corresponding methodological rigour. The systematic review demonstrated that antisocial behaviour instruments reflect multiple generations of psychometric practice, ranging from principal component analysis and sole reliance on Cronbach's alpha to contemporary confirmatory modelling and multi-group invariance testing (Fabrigar et al., 1999; Dunn et al., 2014; Hayes & Coutts, 2020; Brown, 2015; Meredith, 1993). This progression mirrors broader developments within psychological

measurement. However, variability in analytic approach directly affects interpretability and comparability across studies.

Robust evaluation of antisocial behaviour measures follows a cumulative and sequential logic. The empirical evaluation of the ABQ exemplified this staged process, moving systematically from dimensional testing through reliability estimation, invariance analysis, and external validation within a unified analytic framework. Structural validity is foundational, followed by reliability estimation, measurement invariance, and multiple forms of validity evidence (Brown, 2015). This staged approach, increasingly codified within frameworks such as COSMIN, reflects integration of classical test theory and structural equation modelling (Mokkink et al., 2018; Prinsen et al., 2018).

A central issue concerns dimensionality: whether antisocial behaviour is best represented as a general liability or as differentiated components. Dimensional findings are shaped not only by theory but by item architecture and the breadth and depth with which facets are sampled (Clark & Watson, 1995; Reise, 2012). Instruments that sample broadly across behavioural domains using relatively few heterogeneous items often yield a general factor, whereas longer batteries with dense sampling within specific domains are more likely to produce differentiated subscales. Apparent disagreement across measures may therefore reflect differences in sampling depth rather than fundamental construct conflict. Exploratory and confirmatory factor analyses permit formal comparison of single-factor, correlated-factor, and hierarchical representations (Brown, 2015; Reise, 2012). Establishing an appropriate structural model is essential, as reliability and validity coefficients are interpretable only within a supported dimensional framework. In applied contexts where summed scores are common, latent modelling clarifies whether a total score captures a coherent general propensity or conceals domain-specific variance (Reise et al., 2013).

For antisocial behaviour specifically, these issues are substantive. As a developmentally dynamic construct spanning behaviours of varying severity, AB can be conceptualised within a hierarchical framework comprising a general liability toward norm violation alongside specific behavioural expressions (e.g., aggression, deceit, rule violation). This structure parallels broader externalising and "p-factor" models in which shared variance reflects general propensity and domain-specific variance captures differentiated manifestation (Krueger et al., 2005; Caspi et al., 2014). Brief behavioural indices are typically designed to estimate the general liability, whereas extended batteries may differentiate domains when item density permits. Dimensional interpretation must therefore be aligned with the intended level of inference.

Following structural evaluation, reliability indices quantify measurement precision. While alpha remains widely reported, it assumes tau-equivalence and may misestimate reliability when loadings vary (Dunn et al., 2014; McNeish, 2018). Estimators such as McDonald's omega provide more flexible assessment and are increasingly recommended alongside alpha (Hayes & Coutts, 2020). Where relevant, test–retest and inter-rater indices further establish temporal and informant stability. Reliability reflects consistency, not construct accuracy.

Measurement invariance extends structural validity by testing whether the construct operates equivalently across groups or time. Multi-group CFA permits sequential testing of configural, metric, scalar, and residual invariance (Meredith, 1993; Chen, 2007; Putnick & Bornstein, 2016). In developmental research, this is particularly critical. Within the age–crime curve, behavioural prevalence increases in mid-adolescence and declines in early adulthood. Invariance testing clarifies whether observed group differences reflect genuine behavioural change or differential measurement functioning. Without scalar invariance, mean comparisons risk conflating developmental trends with artefact.

The implications differ for brief screeners and extended batteries. For concise indices intended to estimate general antisocial liability, invariance establishes whether a single total score retains comparable meaning across age and gender, supporting risk stratification and longitudinal monitoring. For multi-domain instruments, invariance determines whether both general and subscale factors operate equivalently; failure at the subscale level constrains domain-level interpretation even if a higher-order factor remains stable. Routine invariance testing therefore strengthens defensible developmental comparison.

Construct validity is cumulative. Convergent and divergent validity evidence evaluate whether the measure relates as expected to theoretically neighbouring constructs while remaining distinguishable from conceptually distinct domains (Campbell & Fiske, 1959). More broadly, construct validity is strengthened when findings place the measure coherently within an established nomological network (Cronbach & Meehl, 1955). Criterion validity examines correspondence with diagnostic benchmarks, and predictive validity evaluates temporal forecasting of relevant outcomes.

For measures intended for applied screening, clinical utility must also be evaluated. Receiver operating characteristic (ROC) analyses quantify sensitivity, specificity, and threshold trade-offs (Youden, 1950; Altman & Bland, 1994), particularly relevant in community samples with low base rates of severe conduct problems. ROC procedures translate dimensional evidence into operational decision criteria, linking psychometric robustness to applied implementation.

Within this portfolio, the systematic review examined the extent to which existing antisocial behaviour instruments demonstrated evidence across these sequential stages. The empirical evaluation of the ABQ operationalised this framework directly, testing dimensional structure through latent modelling, estimating reliability beyond alpha, evaluating invariance across gender and developmental stages, and situating the measure within a broader validity and diagnostic framework. Considered together, these components illustrate a structured approach

to antisocial behaviour measurement in which construct definition, item architecture, and sequential validation are aligned. More broadly, they reinforce that the measurability of antisocial behaviour depends not only on theoretical clarity but on transparent and cumulative psychometric evaluation that supports defensible developmental and applied interpretation.

Feasibility, Informant Perspectives, and Applied Use

The systematic review identified a consistent tension between diagnostic breadth and practical feasibility in antisocial behaviour assessment. Instruments demonstrating comprehensive behavioural coverage and strong structural evidence were often lengthy, licensed, or required trained administration. Briefer measures prioritised efficiency and scalability but sampled behavioural domains more sparsely. Instrument selection therefore involves balancing coverage, precision, and administrative burden, a trade-off long recognised in scale development literature (Clark & Watson, 1995; DeVellis, 2016).

Feasibility considerations are particularly relevant in community, school, and large cohort contexts, where time, cost, and respondent fatigue constrain implementation. Extended diagnostic batteries may offer detailed domain-level differentiation but are rarely practical for routine screening or repeated measurement. Brief instruments are more adaptable to large-scale or longitudinal use, provided that structural coherence and measurement stability are established through formal psychometric evaluation (Brown, 2015; Mokkink et al., 2018).

Informant perspective further influences interpretation. Overt behaviours may be reliably observed by caregivers or teachers, whereas covert or peer-context behaviours may be more accessible through youth self-report. No single informant captures the full behavioural range, and discrepancies between informants are common in child and adolescent psychopathology assessment (De Los Reyes et al., 2015). Multi-informant assessment enhances contextual coverage but increases complexity and resource demands.

Clear distinction between screening and diagnostic purposes is therefore essential. Screening tools estimate general liability and support triage or monitoring, whereas diagnostic instruments assess symptom presence, impairment, and contextual specificity in line with formal criteria (American Psychiatric Association, 2013). Use of brief screeners as diagnostic substitutes risks over-interpretation, whereas reliance on extended batteries for population screening may be impractical. Appropriate instrument selection therefore depends on clarity regarding intended level of inference and decision-making context.

Finally, validation context defines interpretive limits. Instruments evaluated in community samples may not generalise directly to forensic or high-risk clinical populations, where prevalence, severity, and base rates differ substantially (Edens et al., 2001). Applied use should therefore align with the population and purpose for which psychometric evidence has been established.

Critical Evaluation of the Portfolio

Strengths

A principal strength of this portfolio lies in its structured integration of construct definition and psychometric evaluation. The systematic review combined a 12-tier taxonomic audit with DSM-5 Conduct Disorder domain alignment (American Psychiatric Association, 2013) and COSMIN-based appraisal (Mokkink et al., 2018), providing a transparent and reproducible method for evaluating antisocial behaviour instruments. The inclusion of 125 unique instrument variants across over 500,000 participants offered a broad evidential base and clarified field-level heterogeneity.

The empirical evaluation extended this foundation by demonstrating that a concise behavioural instrument can meet contemporary standards of structural validity, developmental invariance, and diagnostic evaluation within large community cohorts. Use of

latent modelling (Brown, 2015), multi-group measurement invariance across gender and development (Meredith, 1993; Putnick & Bornstein, 2016), and examination of convergent, predictive, and diagnostic performance provided comprehensive evaluation within two independent UK cohorts. Replication across ROOTS and NSPN enhanced robustness and reduced sample-specific bias. Together, the two studies demonstrate alignment between conceptual mapping and empirical validation.

Limitations

Several limitations warrant consideration. First, the use of DSM-5 Conduct Disorder domains as a comparative anchor represents one operational framework among several possible dimensional models of antisocial behaviour. Alternative hierarchical or spectrum-based models (e.g., Krueger et al., 2005; Caspi et al., 2014) may prioritise partially overlapping constructs and emphasise different structural interpretations.

Second, the systematic review was restricted to English-language, peer-reviewed studies, potentially limiting global representativeness and contributing to over-representation of WEIRD samples (Henrich et al., 2010). Taxonomic classification required interpretative judgement, and COSMIN ratings were necessarily dependent on the quality of reporting within primary studies (Mokkink et al., 2018).

Third, the empirical evaluation relied primarily on community-based samples and youth self-report. Although appropriate for capturing covert behaviours, findings may not generalise directly to forensic or high-risk clinical populations, where prevalence and severity distributions differ (Edens et al., 2001). Sparse endorsement of high-severity behaviours limited modelling of extreme conduct, a recognised challenge in categorical latent variable modelling with low base-rate indicators (Brown, 2015) and may constrain interpretation at the upper tail of the distribution.

Finally, although latent modelling strengthens structural inference, applied contexts frequently rely on summed scores. The translation from latent structure to operational scoring remains an inherent simplification, particularly where essential unidimensionality has not been formally established (Reise et al., 2013).

Methodological Harmony

A defining feature of this portfolio is methodological coherence across its components. The systematic review established explicit criteria for evaluating antisocial behaviour measures, including structural validity, reliability, invariance, and multi-faceted validity evidence (Mokkink et al., 2018). The empirical study subsequently applied these criteria directly to the ABQ within large developmental cohorts using latent modelling and sequential validation procedures (Brown, 2015; Meredith, 1993). This sequential alignment ensures that the empirical evaluation did not operate independently of the review but was explicitly informed by it.

In this respect, the portfolio advances a unified framework for antisocial behaviour measurement, linking construct delimitation, taxonomic clarification, and staged psychometric validation within a coherent methodological programme.

Implications and Future Directions

The findings of this portfolio have implications for both research standards and applied assessment. The systematic review underscores the value of integrating behavioural coverage with contemporary psychometric evaluation, including formal testing of dimensional structure (Brown, 2015), reliability beyond sole reliance on Cronbach's alpha (Dunn et al., 2014; Hayes & Coutts, 2020), and routine examination of measurement invariance in developmental samples (Meredith, 1993; Putnick & Bornstein, 2016). Transparent reporting

of model fit indices and invariance testing would strengthen comparability across studies and reduce interpretive ambiguity (Brown, 2015).

Future research should extend these standards across more diverse populations. Cross-national invariance testing is necessary to determine whether structural stability generalises across cultural contexts (Putnick & Bornstein, 2016), and validation within clinical and forensic samples is required to evaluate performance under differing base rates and severity distributions (Edens et al., 2001). Multi-informant evaluation of brief behavioural indices would clarify contextual visibility of behaviours and inform interpretation across settings (De Los Reyes et al., 2015). Additionally, sensitivity-to-change analyses are needed to establish whether brief dimensional indices are responsive to intervention or developmental transition, a prerequisite for monitoring use in applied contexts (Mokkink et al., 2018).

For clinical and educational settings, the findings emphasise the importance of aligning instrument selection with purpose. Brief screeners may support early identification and longitudinal monitoring where feasibility constraints are present, whereas comprehensive batteries remain necessary for diagnostic clarification in accordance with formal criteria (American Psychiatric Association, 2013). Interpretation of developmental or gender differences should remain cautious in the absence of established measurement invariance (Meredith, 1993; Chen, 2007). Applied use should therefore remain anchored to the evidential context in which psychometric properties have been established.

Conclusion

This portfolio addressed the conceptual and methodological coherence of adolescent antisocial behaviour measurement through two complementary approaches. The systematic review clarified the heterogeneity of existing instruments and established structured criteria integrating behavioural coverage with contemporary psychometric evaluation (Brown, 2015;

Mokkink et al., 2018). The empirical evaluation applied these standards directly to a brief DSM-aligned behavioural index, demonstrating how dimensional structure, reliability, measurement invariance (Meredith, 1993; Putnick & Bornstein, 2016), and validity evidence can be examined within large developmental cohorts. Together, the studies reinforce that defensible measurement of antisocial behaviour requires alignment between construct delimitation, item architecture, and sequential validation procedures.

At a field level, this integration supports more transparent instrument selection, improved comparability across developmental and demographic groups, and clearer distinction between screening and diagnostic purposes. By situating behavioural specificity alongside psychometric rigour and applied feasibility, the portfolio contributes a coherent framework for evaluating and implementing antisocial behaviour measures within research and community contexts.

References

Altman, D. G., & Bland, J. M. (1994). Diagnostic tests 3: Receiver operating characteristic plots. BMJ, 309(6948), 188. https://doi.org/10.1136/bmj.309.6948.188

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (5th ed.). American Psychiatric Publishing.

Brown, T. A. (2015). Confirmatory factor analysis for applied research (2nd ed.). Guilford Press.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. Psychological Bulletin, 56(2), 81–105. https://doi.org/10.1037/h0046016

Caspi, A., Houts, R. M., Belsky, D. W., Goldman-Mellor, S. J., Harrington, H., Israel, S., Meier, M. H., Ramrakha, S., Shalev, I., Poulton, R., & Moffitt, T. E. (2014). The p factor: One general psychopathology factor in the structure of psychiatric disorders? Clinical Psychological Science, 2(2), 119–137. https://doi.org/10.1177/2167702613497473

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. Structural Equation Modeling, 14(3), 464–504. https://doi.org/10.1080/10705510701301834

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. Psychological Assessment, 7(3), 309–319. https://doi.org/10.1037/1040-3590.7.3.309

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52(4), 281–302. https://doi.org/10.1037/h0040957

De Los Reyes, A., Augenstein, T. M., Wang, M., Thomas, S. A., Drabick, D. A. G., Burgers, D. E., & Rabinowitz, J. (2015). The validity of the multi-informant approach to assessing child and adolescent mental health. Psychological Bulletin, 141(4), 858–900. https://doi.org/10.1037/a0038498

DeVellis, R. F. (2016). Scale development: Theory and applications (4th ed.). Sage.

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. British Journal of Psychology, 105(3), 399–412. https://doi.org/10.1111/bjop.12046

Edens, J. F., Skeem, J. L., Cruise, K. R., & Cauffman, E. (2001). Assessment of "juvenile psychopathy" and its association with violence. Behavioral Sciences & the Law, 19(1), 53–80. https://doi.org/10.1002/bsl.425

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. Psychological Methods, 4(3), 272–299. https://doi.org/10.1037/1082-989X.4.3.272

Frick, P. J., Ray, J. V., Thornton, L. C., & Kahn, R. E. (2014). Annual research review: A developmental psychopathology approach to understanding callous–unemotional traits. Journal of Child Psychology and Psychiatry, 55(6), 532–548. https://doi.org/10.1111/jcpp.12152

Hare, R. D., & Neumann, C. S. (2008). Psychopathy as a clinical and empirical construct. Annual Review of Clinical Psychology, 4, 217–246. https://doi.org/10.1146/annurev.clinpsy.3.022806.091452

Hayes, A. F., & Coutts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. Communication Methods and Measures, 14(1), 1–24. https://doi.org/10.1080/19312458.2020.1718629

Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content validity in psychological assessment. Psychological Assessment, 7(3), 238–247. https://doi.org/10.1037/1040-3590.7.3.238

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? Behavioral and Brain Sciences, 33(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Krueger, R. F., & Markon, K. E. (2006). Reinterpreting comorbidity: A model-based approach. Annual Review of Clinical Psychology, 2, 111–133. https://doi.org/10.1146/annurev.clinpsy.2.022305.095213

Krueger, R. F., Markon, K. E., Patrick, C. J., Benning, S. D., & Kramer, M. D. (2005). Linking antisocial behavior, substance use, and personality. Journal of Abnormal Psychology, 114(4), 537–548. https://doi.org/10.1037/0021-843X.114.4.537

Lahey, B. B., Applegate, B., Hakes, J. K., Zald, D. H., Hariri, A. R., & Rathouz, P. J. (2012). Is there a general factor of prevalent psychopathology? Journal of Abnormal Psychology, 121(4), 971–977. https://doi.org/10.1037/a0028355

Marsh, H. W. (1994). Sport motivation orientations: Beware of jingle–jangle fallacies. Journal of Sport & Exercise Psychology, 16(4), 365–380. https://doi.org/10.1123/jsep.16.4.365

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. Psychological Methods, 23(3), 412–433. https://doi.org/10.1037/met0000144

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. Psychometrika, 58(4), 525–543. https://doi.org/10.1007/BF02294825

Moffitt, T. E., Caspi, A., Harrington, H., & Milne, B. J. (2002). Males on the life-course-persistent and adolescence-limited antisocial pathways. Development and Psychopathology, 14(1), 179–207. https://doi.org/10.1017/S0954579402001104

Mokkink, L. B., Prinsen, C. A. C., Patrick, D. L., Alonso, J., Bouter, L. M., de Vet, H. C. W., & Terwee, C. B. (2018). COSMIN methodology for systematic reviews of PROMs. Quality of Life Research, 27(5), 1171–1179. https://doi.org/10.1007/s11136-018-1798-3

Prinsen, C. A. C., Mokkink, L. B., Bouter, L. M., Alonso, J., Patrick, D. L., de Vet, H. C. W., & Terwee, C. B. (2018). COSMIN guideline for systematic reviews of PROMs. Quality of Life Research, 27(5), 1147–1157. https://doi.org/10.1007/s11136-018-1798-3

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting. Developmental Review, 41, 71–90. https://doi.org/10.1016/j.dr.2016.06.004

Reise, S. P. (2012). The rediscovery of bifactor measurement models. Multivariate Behavioral Research, 47(5), 667–696. https://doi.org/10.1080/00273171.2012.715555

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modelling psychological measures in the presence of multidimensionality. Journal of Personality Assessment, 95(2), 129–140. https://doi.org/10.1080/00223891.2012.725437

Sampson, R. J., & Laub, J. H. (1993). Crime in the making: Pathways and turning points through life. Harvard University Press.

Shulman, E. P., Smith, A. R., Silva, K., Icenogle, G., Duell, N., Chein, J., & Steinberg, L. (2016). The dual systems model. Psychological Science, 27(3), 377–389. https://doi.org/10.1177/0956797615624771

Steinberg, L. (2008). A social neuroscience perspective on adolescent risk-taking. Developmental Review, 28(1), 78–106. https://doi.org/10.1016/j.dr.2007.08.002

Youden, W. J. (1950). Index for rating diagnostic tests. Cancer, 3(1), 32–35. https://doi.org/10.1002/1097-0142(1950)3:132::AID-CNCR2820030106>3.0.CO;2-3

Appendices

Appendix A: Confirmation of Full Ethical Approval from FMH REC

University of East Anglia
Norwich Research Park
Norwich. NR4 7TJ

Email: ethicsmonitor@uea.ac.uk
Web: www.uea.ac.uk

Study title: The Antisocial Behaviours Questionnaire (ABQ): A Systematic Evaluation of Psychometric Properties using two large-scale UK samples.

Application ID: ETH2425-2404

Dear Punita,

Your application was considered on 25th June 2025 by the FMH S-REC (Faculty of Medicine and Health Sciences Research Ethics Subcommittee).

The decision is: **approved**.

You are therefore able to start your project subject to any other necessary approvals being given.

This approval will expire on **30th September 2026**.

Please note that your project is granted ethics approval only for the length of time identified above. Any extension to a project must obtain ethics approval by the FMH S-REC (Faculty of Medicine and Health Sciences Research Ethics Subcommittee) before continuing.

It is a requirement of this ethics approval that you should report any adverse events which occur during your project to the FMH S-REC (Faculty of Medicine and Health Sciences Research Ethics Subcommittee) as soon as possible. An adverse event is one which was not anticipated in the research design, and which could potentially cause risk or harm to the participants or the researcher, or which reveals potential risks in the treatment under evaluation. For research involving animals, it may be the unintended death of an animal after trapping or carrying out a procedure.

Any amendments to your submitted project in terms of design, sample, data collection, focus etc. should be notified to the FMH S-REC (Faculty of Medicine and Health Sciences Research Ethics Subcommittee) in advance to ensure ethical compliance. If the amendments are substantial a new application may be required.

Approval by the FMH S-REC (Faculty of Medicine and Health Sciences Research Ethics Subcommittee) should not be taken as evidence that your study is compliant with the UK General Data Protection Regulation (UK GDPR) and the Data Protection Act 2018. If you need guidance on how to make your study UK GDPR compliant, please contact the UEA Data Protection Officer (dataprotection@uea.ac.uk).

Please can you send your report once your project is completed to the FMH S-REC (fmh.ethics@uea.ac.uk).

I would like to wish you every success with your project.

On behalf of the FMH S-REC (Faculty of Medicine and Health Sciences Research Ethics Subcommittee)

Yours sincerely,

Dr Paul Linsley