Article

Crime & Delinquency
1–29
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/00111287231180096
journals.sagepub.com/home/cad
Sage

Best Practices for Neglected Assumptions: Multi-Site Confirmation of the MPACT-6

Xiaohan Mei[1], Zachary Hamilton[2], Alex Kigerl[2], Amber Krushas[2], and Faye S. Taxman[3]

Abstract

In the assessment world, risk determines "who" to treat and needs determine "what" (Bonta & Andrews, 2016). Yet, for youth, greater emphasis is placed on needs that target recidivism reduction interventions. Unlike risk, needs represent dynamic domains, or latent constructs, requiring testing to assure adequate measurement. We conducted a multi-level, multi-group analysis of the Modified Positive Achievement Change Tool (MPACT) with a 10-state sample of youth (*N*=258,464). Findings confirm the validity and reliability of needs domains, and the development of a novel "Global Needs Factor," a composite summary of needs. Created without criminal history/static measures, needs scales demonstrate predictive accuracy. Further, measurement invariance and aspects of gender and race/ethnicity prediction parity are observed.

Keywords

risk-needs-assessment, juvenile justice, overclassification, measurement invariance

[1]California State University – Los Angeles, USA
[2]University of Nebraska – Omaha, USA
[3]George Mason University, Arlington, VA, USA

*Amber Krushas is now affiliated with University of Nevada – Las Vegas, Department of Criminal Justice.

Corresponding Author:
Xiaohan Mei, School of Criminal Justice and Criminalistics, California State University – Los Angeles, Hertzberg-Davis Forensic Science Center, 5151 State University Dr., Los Angeles, CA 90032, USA.
Email: xmei4@calstatela.edu

Introduction

Grounded in D. A. Andrews and Bonta's (2010) risk, need, and responsivity (RNR) principles, Risk-Needs Assessments (RNAs) allow justice professionals to collect necessary information to classify individuals' likelihood of recidivism and factors to be addressed through programing. Over the past four decades, RNA utilization has expanded to nearly every state and justice population (Juvenile Justice Geography, Policy, Practice & Statistics, 2020). While static "risk" is conceptualized as prior history and justice system involvement, dynamic risk or "needs" represent dynamic features that are functionally changeable and amendable to services and interventions (D. A. Andrews & Bonta, 2010). Thus, RNA tools can be used to guide individuals' level of supervision and programing. The RNR model advocates for using supervision and correctional programing to reduce recidivism, outlining that risk identifies "who" to treat, and needs identify "what" to treat (Bonta & Andrews, 2016). Specifically, a person's risk score provides a comprehensive summary of RNA items. Yet, risk scores include criminal history and other static indicators, thus preventing a summary needs assessment. For youth, criminal history and other static indicators are less prevalent and predictive, leading to calls for the development of a measure that will summarize programmatic targets of youth (Taxman & Smith, 2021).

Parallel to adult tools, juvenile RNAs (JRNAs) have similar designs and content. Yet, a "cultural shift in juvenile justice" supports greater emphasis on evidence-based practices, recognizing youths' potential strengths, needs, and safety (Vincent et al., 2012, p. 18). This has propelled an even greater emphasis on reducing youth needs as a tool to reduce justice system contact. Further, foundational correctional theory outlines the use of needs to prioritize programing and placement (Bonta & Andrews, 2016). Yet, contemporary tools lack the ability to classify and summarize needs. While several contemporary JRNAs exist, most have been modified from adult tools, with validation efforts focused on establishing accuracy of static risk scales in predicting recidivism (J. L. Skeem et al., 2013; Taxman, 2017).

Dynamic needs are often built around the same common eight factors (Jung & Rawana, 1999; Wormith & Bonta, 2018) with similar items and content theorized as constructs based on less-than-optimal development (see D. A. Andrews & Bonta, 2010). While many contemporary risk tools have demonstrated predictive validity, the industry standard bar for this claim is easily achieved (Hamilton et al., 2017). To advance these theoretical constructs, greater evidence must be provided to establish both the existence and appropriate use of needs domains (Taxman & Caudy, 2015). In other words, while theoretically associated with recidivism, empirical evidence is needed to support needs domains both in terms of construct validity (Rios & Wells, 2014) and predictive validity. According to Bonta (2002), all construct validity components (i.e., factor structure, reliability, convergent validity, discriminative validity, face validity, predictive validity) are, "important psychometric characteristics" of needs scales (p.358). However, the validity components of needs domains are rarely examined (Taxman, 2017).

Without construct validity, developers cannot confirm their domains represent needs, which may lead to over/underestimating (J. Skeem et al., 2016) and/or a misallocation limited resource (Taxman, 2017). Further validation evidence should lead to new versions, connecting needs and interventions to address youth needs and inform practice (J. P. Singh et al., 2014).

Concerns have also been raised regarding RNAs' ability to provide equivalent assessment and proper classification across gender and race/ethnicity. Specifically, overclassification has been identified within RNAs where certain subgroups (e.g., females, minorities) have the same risk score as other groups, yet recidivate at lower rates (Angwin et al., 2016; Hamilton et al., 2019). RNAs' overreliance on static indicators contribute to overclassification (W. T. Miller et al., 2022), where greater reliance on needs items may create greater prediction parity (Butler et al., 2023).

Theoretical constructs should demonstrate evidence across multiple populations of varied patterns. To date, no tools have empirically confirmed the efficacy of needs scales across sites, system stages, or key sub-groups (e.g., gender, race). Extending prior work on the Modified Positive Achievement Change Tool (MPACT), the current study examined the risk and need domains' internal structure and creation of a "Global Needs Factor" using a 10-state, U.S. representative sample of justice involved youth (N=258,464 youth). Through confirmation of MPACT needs scales, we extend examinations beyond those of prior tools, presenting evidence of gender and race invariance. Further, we assess predictive validity, evaluating parity across gender and race lines. This study then establishes a generally applicable needs tool for youth.

Literature Review

The modern RNA is thought to have begun with the Level of Service suite of tools (D. A. Andrews & Bonta, 1995). The Level of Service/Case Management Inventory and its youth version (LS/CMI & YLS/CMI, respectively) were designed to measure risk, need, and programing responsivity. Developed as the "Central Eight," the tool is made of eight domains, seven of which conceptualized to measure needs, including (1) Substance Use, (2) Antisocial Personality, (3) Antisocial Cognitions, (4) Antisocial Associates, (5) Family and Marital Relations, (6) Employment, and (7) Leisure and Recreational Activities (D. A. Andrews & Bonta, 2010). As outlined in the RNR model, to be included in a needs domain, items must be dynamic (i.e., can change over time) and associated with recidivism. The developers created a pool of hundreds assessment items following a "brainstorming session" with Canadian probation officers. In subsequent meetings, they reduced the number of items and piloted the tool with 112 halfway house participants. After multiple tests within Ontario and Manitoba, the LS/CMI was created, which rearranged the items to form the Central Eight and align with the General Personality and Cognitive Social Learning (GPCSL) theory (Wormith & Bonta, 2018).[1] Furthermore, because needs change over time, constructed domains represent the target of correctional programing. The LS tools are comprised of 54 unweighted Burgess items (0/1) that are summed across all eight domains to create a composite score, where predictive validity of the youth tool ranges widely from small-to-large (AUCs = 0.57–0.75) and are notably weaker for US samples (Olver et al., 2014).

Another commonly used contemporary tool, the Ohio Risk Assessment System and its youth counterpart (ORAS, OYAS) were created in 2010. Modeled after the LS/CMI tools, risk and needs domains provide similar Central Eight-like content, albeit with fewer scoring items, which present shorter need domain ranges. While needs domains provide sub-scores indicating programing targets, the Ohio tools also provide an unweighted composite risk score that demonstrate moderate predictive accuracy (AUCs = 0.64–0.69).

The Positive Achievement Change Tool (PACT) and Youth Assessment Screening Inventory (YASI) are two assessments derived from the same tool, developed from a sample of Washington State youth on probation (Barnoski, 1997). The tool uses a scoring algorithm of regression weighted items, where an abbreviated version (pre-screen) is used to assess risk of recidivism, demonstrating moderate predictive accuracy (AUC = 0.56–0.70). Youth scoring moderate or high-risk receive a more extensive, 10-domain[2] needs assessment.[3]

Construct Validity of Need Domains

As mentioned, while predictive validity is often assessed for an instrument's risk score, critical tests of needs assessments' construct validity are often neglected. For instance, among their review of 16 youth risk assessments, Onifade et al. (2009) found only three tools were validated more than once, and most examined predictive validity exclusively. Notably, needs domains represent clusters of items that collectively represent a "latent construct," where each item is indirectly associated with the manifest outcome—recidivism—and other items in the domain through their relationship to the construct. Because needs domains represent latent measures, more extensive construct validity testing is required to ensure the content of a given domain measures the need in question. Some of the most important construct validity assessments include face, convergent and divergent, internal structure, and predictive validity.

Briefly, face validity is the extent to which users perceive the domain's item content to have convergence and relevance for individuals in which the tool is administered, while convergent and divergent validity assesses if domain items are correlated with each other and not with items in other domains (Hsu et al., 2010).[4] The construct validity assessment of this paper focuses on the MPACT's structural validity, which includes dimensionality, reliability, and measurement invariance, while also assessing its predictive validity. Specifically, dimensionality examines the hypothesized inter-relationships between needs items and the latent variables. Using prior theory and evidence as a guide, confirmatory factor analyses (CFA) is used to assess if item loadings identify a singular domain or multiple sub-domains/dimensions (Rios & Wells, 2014). Next, reliability indices assess the proportionality of the true score variance to the total score variance (Rios & Wells, 2014). Measurement invariance assesses if domain items are scored similarly across key groups (e.g., gender, race). Finally, predictive validity assesses if domains, and the larger summary score, predict the manifest outcome—recidivism.[5]

While many of the most widely used assessments are constructed of both risk and needs tools, agencies often struggle to fully utilize assessments and are limited in applying program options to identified needs (Taxman, 2017). The foundation of a needs tool begins with face validity, where needs domains provide "relevant" content to reduce youth deficits (or enhance strengths) and, we argue, should tie to common correctional interventions. All contemporary RNAs acknowledge the Central Eight as a foundation of needs assessment (see Scott et al., 2019). Yet, practitioners have expressed face validity concerns, as constructs are not well-described (e.g., antisocial personality) and it is difficult to link interventions to all domains (i.e., leisure and recreation) (J. Miller & Maloney, 2013; Taxman & Caudy, 2015). Thus, practitioners may question the appropriateness and utility of needs assessment results and have been known to complete the tool, file it away, and ignore recommendations (Viglione et al., 2015).

The Central Eight's development from a 1980s era "brainstorming session" should not be overlooked, as empirical findings have not established their existence as needs constructs. To date, tests of the Central Eight's internal structure validity only exist for adult tools (e.g., D. A. Andrews & Bonta, 1995; Hollin et al., 2003; Jung & Rawana, 1999; Wormith & Bonta, 2018), and notably, findings do not meet psychometric industry standards (American Educational Research Association, American Psychological Association, National Council on Measurement in Education (2014)). Specifically, studies cannot confirm that all domains possess dimensionality (e.g., D. A. Andrews et al., 2006; Palmer & Hollin, 2007; N. Schmidt et al., 2017; Stevenson & Wormith, 1987), reliability (F. Schmidt et al., 2005), or measurement invariance across sex/gender (Kitzmiller et al., 2022). Further, inconsistent support for the predictive validity of the Central Eight has been found[6] by comparison to Canadian and White samples (Olver et al., 2014). Despite a lack of empirical support for adult populations, the Central Eight were still applied to juvenile populations via the YLS/CMI and to date, no tests of internal structural validity have been completed.

JRNAs generally have received scant validation and limited updates (F. Schmidt et al., 2005). Sullivan et al. (2019) examined a three-state sample of OYAS-assessed youth, identifying small predictive effects for needs domains (AUC = 0.51–0.63). Then, in 2021, Sullivan and colleagues examined the OYAS' internal structure, where findings could not confirm OYAS domains. Further, findings indicated inconsistent recidivism prediction for OYAS needs across all races/ethnicities. Regarding the PACT/YASI, Barnoski (2004a) provided a needs tool analysis, which failed to confirm the construct validity of domains.

Gender and Race Parity

Many tools make claims of gender-and race-neutrality, based on predictive validity findings a tool's risk score. Specifically, prior work has found the predictive validity metrics (i.e., correlation coefficient, AUCs) of the PACT, OYAS, and YLS/CMI are mixed to relatively equal for white and non-white individuals (e.g., Baglivio & Jackowski, 2013; Childs et al., 2022; Jung & Rawana, 1999; Olver et al., 2014; Vincent et al., 2011). Other work has also found these metrics to be relatively equal for males and females (e.g., Childs et al., 2022; Olver et al., 2014).

Nevertheless, concerns have become pronounced due to recent findings of overclassification for race-and gender-neutral tools (e.g., C. A. Campbell et al., 2020; W. T. Miller et al., 2022; Zottola et al., 2022). For instance, in a seminal piece examining the COMPAS, ProPublica identified a greater proportion of minority individuals as High-Risk, yet a lower rate of recidivism was observed for High-Risk minorities when compared to their White counterparts (i.e., false positive rate [FPR]) (Angwin et al., 2016). Further, Hamilton et al. (2019) examined the predictive parity of the PACT, identifying a lower recidivism rate of High-Risk females, as compared to males. Additionally, studies demonstrate a lack of measurement invariance across race and sex sub-groups among various youth tools (e.g., YLS/CMI, OYAS; Kitzmiller et al., 2022; Sullivan et al., 2022), leading to questions of tool legitimacy (S. Schmidt et al., 2020).

Critiques have suggested the use of criminal history and other static measures may cause differential assessment across race and gender. As it relates to race, W. T. Miller et al. (2022) indicated that minorities are more likely to live in areas with greater police presence, and thus, RNA items that pertain to static and justice system contacts are disproportionate and may result in overclassification. Further, W. T. Miller et al. (2022) identified that when added to dynamic scales, LS/CMI static predictors only improved prediction for White youth. Notably, life-course research has identified some delinquency is normal in adolescence (Henning, 2012), questioning the need to heavily weigh criminal history items for youth.

For females, Van Voorhis et al. (2010) argued that RNAs are developed with a substantially greater proportion of males, and when items are selected and weighted, they more often represent male risk. Therefore, when applied to females, criminal history and other static indicators that are predictive for males often lead to overclassification for females. In fact, Belisle and Salisbury (2021) reviewed five popular juvenile RNA tools across 21 empirical studies and identified that overclassification occurred in 74% of the examinations.

Recent work identifies needs tools as a promising avenue for future development. For instance, using two large juvenile samples, Butler et al. (2023) and Hamilton et al. (2019) identified that optimizing tools to include fewer criminal history items and greater needs items ameliorate biases and overclassification with only a small sacrifice to predictive accuracy. Additionally, Wong and Gordon (2006) identified that the dynamic items of the Violence Risk Scale (VRS) were as predicative as their entire tool. Unfortunately, comparisons between key demographic groups cannot be made without establishing invariance, as there is no guarantee a domain has the same effect for different genders and races (Pardoel, 2020). Thus, while some RNA tools claimed to be gender- and race-neutral, the empirical evidence suggests otherwise. Without support that needs assessments are unbiased, minority and female groups will be overclassified and receive unnecessary programing and supervision (Taxman & Smith, 2021). Given the field's emphasis on the dynamic needs of youth (Caudy et al., 2013), we sought the development of a tool, absent static indicators, that provides invariance and predictive parity.

Global Needs Assessment

As indicated, a theoretical postulate of GPCSL and the RNR model is that an individual's risk tells us "who" to treat, while needs domain scores tells us "what" to treat (Bonta & Andrews, 2016). Essentially, prioritizing high-risk for programing and higher scores in needs domains outline the intervention needed. However, as risk scores include criminal history and other static indicators that cannot be altered via programing, this may lead to inappropriate prioritization of programing around historical behaviors. Recently, National Academies of Sciences, Engineering, and Medicine (2022) has raised concerns, as criminal history indicators use administrative measures of system involvement (e.g., arrests, adjudications), rather than deviant or problematic behavior. Thus, targeting high risk, rather than high need, is an inefficient way to reduce future deviant behavior.

However, the prioritization of programing through risk scores may be an artifact of contemporary tool development. Yet, assessments from other fields provide examples of an overall evaluation of needs. The Global Assessment of Functioning (GAF) Scale, for instance, is used by mental health professionals to describe an individual's level of functioning and determine patient interventions (Pedersen et al., 2007). Historically, RNAs only provide a total summary score of risk. Unfortunately, contemporary RNAs do not provide a global needs score, leading to calls for development (Taxman & Smith, 2021).

The Development of the Modified Positive Achievement Tool (MPACT)

The Washington State Juvenile Court Assessment—Risk Assessment (WSJCA-RA), was developed to guide youth supervision, programing, and case planning (Barnoski, 1997). Given the non-proprietary nature of the tool, over 20 states adopted the tool, leading to multiple name applications (e.g., Positive Achieve Change Tool [PACT], Youth Assessment Screening Inventory [YASI]). Like most tools, need domains were developed from a review of empirical findings, including 10 theoretically derived needs scales[7] (Barnoski, 2004a). Risk models were recently optimized to improve risk prediction (AUC = 0.68–0.84) (Hamilton et al., 2022).

Studies examining the tool's dimensionality indicated a redesign of the needs domain structure was appropriate (Barnoski, 2004b; J. L. Skeem et al., 2013). In 2021, the PACT was updated to the Modified-Achievement Change Tool (MPACT), developing six hypothesized needs domains more closely tied with programing needs of youth and meeting psychometric standards for latent scales (Mei et al., 2021). For this "proof of concept," authors examined redesigned domains using Exploratory Factor Analysis (EFA) for a Washington State probation youth sample. Findings demonstrated convergent and divergent validity of six constructs and improved tool performance, creating a foundation for the larger work presented here (Mei et al., 2021).

Current Study

The goal of this study is to empirically validate the MPACT needs tool and domains using data from 10 states, a variety of justice stages, and by key sub-population (e.g., race, gender). We further developed a "Global Needs Factor" representing a summary score of youth needs. Finally, the predictive validity of each domain scale was tested and compared to assess the tool's predictive parity. In doing so, the current study uses "best practices" for assessing needs tools, attempting to meet four important criteria. First, a needs tool must be composed of only dynamic items, presenting an ability to measure change upon reassessment and provide relevance for program matching. Second, domain structure must be identified and confirmed with psychometric evidence. Third, to warrant general application, reliability of needs domains must be established using multiple youth populations (e.g., sites) and supervision stages (i.e., diversion, probation, parole). Finally, developers must demonstrate that needs domains measure similarly and provide equitable prediction across gender and race/ethnicity sub-groups.

Methods

Collected as part of a larger Office of Juvenile Justice and Delinquency Prevention (OJJDP) study, we obtained 258,464 records from 10-states[8] of justice-involved youths at different justice stages, including diversion, detention as well as probation and parole.[9] The sample featured 30.1% females and was 55.9% White, 34.9% Black, 5.6% Hispanic, and 3.6% "Other" race/ethnicity youth (see Krushas et al., 2023 for additional sample descriptives).

Measures

The WSJCA-RA was used for the current study, which consisting of 132 total items (see Barnoski, 2004b),[10] which were further reduced to 81 dynamic items used to create assessment domains. Responses were collected via structured interviews[11] with youth and their family. When developing constructs, we only utilized the tool's dynamic items. Slight variation in data collection process and assessment formulation was observed and addressed via a reconciliation process by employing traditional data cleaning techniques, such as collapsing and adjusting responses. Furthermore, missing data were addressed using random forest imputation approach.[12] Reader should refer to a report (Krushas et al., 2023) for item response frequencies, descriptive statics, and item loadings. Recidivism represents a new charge that resulted in adjudication within 365 days of a youth's initial assessment, or their community supervision start date.[13]

Analytic Strategy

We first conducted EFAs on domain items to assess dimensionality of the six MPACT constructs.[14] Next, Multi-Group Confirmatory Factor Analyses (MGCFAs) were computed, including higher-order tests and gender and race invariance[15] (Putnick & Bornstein, 2016). Latent constructs were then weighted and combined to create a "Global Needs Factor" ("G-factor"), representing overall level of youth needs. The "G-factor" was also tested for measurement invariance.[16] Construct reliability was assessed using omega coefficient (ω).[17] We evaluated EFA and CFA models comparing model fit indices and loadings/cross-loadings with industry-standard thresholds (Tabachnick & Fidell, 2007).[18] EFA domain factor loadings are provided in a research brief (Krushas et al., 2023).[19] Model fit was also assessed using Comparative Fit Index (CFI)/Tucker Lewis Index (TLI) and the Root Mean Square Error of Approximation (RMSEA).[20] Models were evaluated with constraints added in each additional and progressive model for higher-order and group invariance tests. Higher-order models and those with additional measurement invariance constraints were retained if ΔCFI and ΔTLI values were acceptable (<0.01), indicating models did not detrimentally impact fit (Little, 2013).[21]

Finally, we examined predictive validity of the constructs and the global scale using on recidivism. Predictive discrimination is measured via the AUC statistic (J. Singh, 2013) and represent an effect size, where 0.55 indicates a small effect, 0.63 a moderate effect, and 0.70 a large effect (Rice & Harris, 2005). Next, as a combined measure of discrimination, accuracy, and calibration, the Squared error, Accuracy, and Receiver operating characteristic (SAR) score was computed. These values were broken down for race/ethnicity (White, Black, Hispanic, "Other") and gender (male/female). Finally, following prior overclassification assessments (Angwin et al., 2016; Hamilton et al., 2019), we evaluate predictive parity both graphically and statistically, comparing the global needs score by sub-group. Graphically, we use scatter plot fit lines to trace the pattern of recidivism probability of needs score by sub-group. Then, using the sample's recidivism base rate (48.8%) as a reference, FPRs identify the rate higher-need subjects do not recidivate, while the positive predictive values (PPVs) assess the proportion of higher-need persons who reoffended (J. Singh, 2013). We use the burgeoning industry standard "k-fold validation procedure" to compute predictive validity metrics (Steyerberg et al., 2003).[22]

Results

The results are presented in three sections. First, the design of each developed domain is described, outlining face validity regarding domain item depth available to assess variability in youth needs and functional elements in evaluating a continuum of both needs and protective factors.[23] Next, findings of MGCFA models' dimensionality, reliability, and measurement invariance are presented. Finally, we assess predictive validity and parity.

Domain Design

"Education" is a single-order factor that assesses the extent a youth has attained their educational goals. This domain is designed to provide sufficient item depth ($k=10$) to balance both needs to be addressed (e.g., conduct, attendance) and protective items to be strengthened (e.g., activities involvement, education value).

The "Association" domain assesses the strength of youth commitment to prosocial activities via a third-order factor. This domain combines interrelated content areas of youth "Free Time" and "Pro-Social Attachments" (Commitment), with "Employment" and "Anti-Social Associations" subdomains. This domain provides sufficient item depth ($k=9$) to balance both needs to be addressed (e.g., admires/emulates anti-social peers) and protective items to be strengthened (e.g., prosocial community ties). This domain combines three original PACT/YASI domains (see Barnoski, 2004a).

The "Family" domain is a single-order factor that assesses positive family relationships and supportive environment. Notably, this domain provides a deep ($k=16$) continuum of needs (e.g., run-away) and protective (e.g., family member relationships) items. Next, the "Alcohol & Drugs" domain represents a single-order factor designed to assess youths' overall substance abuse problems and disruptive consequences, proving a deep, 16-item continuum of needs (e.g., use disrupting education) and protective (e.g., drug/alcohol treatment) items.[24] The "Mental Health" domain is also a single-order domain that assesses the extent of youths' recent symptoms and issues. Creating a shorter continuum than other MPACT domains, this construct contains four items,[25] where responses assess needs (e.g., issues interfere with work) and protective items (e.g., compliance with medication).

The sixth domain—"Cognition & Behaviors"—is a third-order domain and assesses the extent youths have internalized pro-criminal definitions and attitudes. Combining three sub-scales (e.g., "Attitude/Behavior," "Aggression," "Skills"), this domain provides an expansive item depth ($k=25$). Reader readers should refer to a research report for items and loadings, along with figures, for each domain (Krushas et al., 2023).

Finally, the Global Needs Factor or "'G-factor" was extracted, representing a weighted composite score of all six domains. Where some individuals may present high needs in one or two areas, prior findings identify that some justice-involved and cross-over youth present high needs across several domains (Herz et al., 2012; Kapoor et al., 2018). This global needs score provides a metric to identify those youth with the greatest intervention needs, where agencies are instructed to use the "G-factor" scores to determine eligibility and prioritization when programing slots are limited. A visual illustration of the G-factor', and the standardized factor (domain) loadings are provided in Figure 1.

Structural Validity

All six constructs and the "G-factor" "passed" gender- and race- invariance tests, where all CFA findings exceeded model fit (CFI & TFI > 0.90, RMSEA > 0.05) and reliability thresholds for unidimensional ($\omega > 0.80$) and multi-dimensional ($\omega > 0.65$). Summary statistics are provided in Table 1.[26] Minor model inconsistencies were identified, where one item[27] loading in the "Education" and five items[28] in the "Family" domain did not exceed model fit thresholds (see Krushas et al., 2023). Nevertheless, these items represent theoretically important content and were thus retained as, model fit statistics exceeded established thresholds, indicating their inclusion did not adversely impact domain measurement. The "Mental Health" and "Alcohol & Drugs" domains demonstrated reduced loading strength by comparison to other MPACT domains. Again, the inclusion of all six constructs provided acceptable model fit and exceeded invariance test thresholds.

Predictive Validity

Due to the large sample size, all AUCs analyses were significant ($p < .001$). "Cognition & Behavior," "Education," and "Associations" domains possessed the greatest prediction strength (AUC=0.65, 0.64, & 0.59, respectively). The "G-factor" indicated moderate prediction strength and

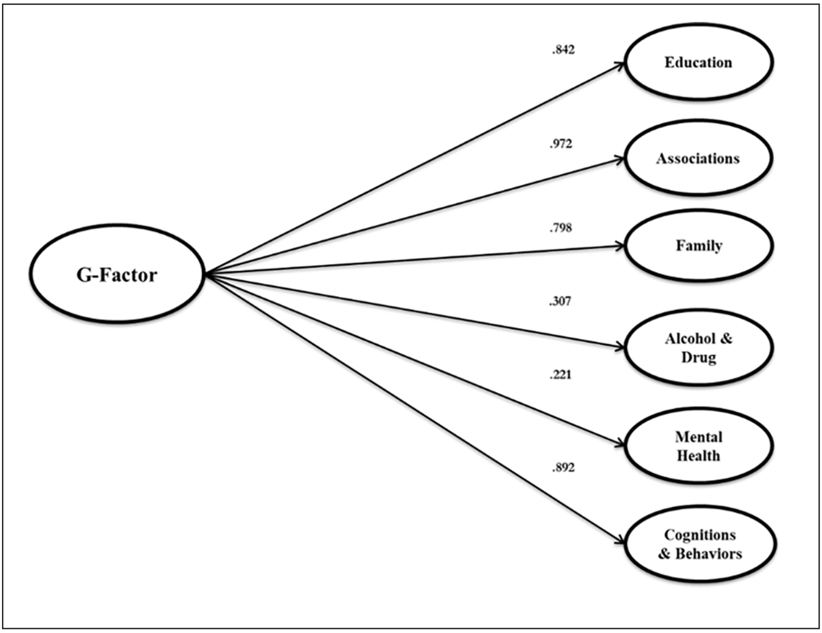

Figure 1. Youths' global Needs Factor (G-Factor).

(AUC = 0.68), presenting prediction strength similar to contemporary JRNA risk scores.

We then examined prior RNA criteria for gender-neutrality comparing AUC differences across gender/sex and race/ethnicity sub-groups. As shown in Table 2, comparatively, differences were all minor, representing less than an effect size range of 0.07 difference between groups.[29] Black youth possessed the lowest AUCs for individual domains, but the "G-factor" was the same as White youth (0.67). Apart from "Family" and "Associations," Hispanic youth possessed the best domain and overall AUC values ("G-factor" AUC = 0.71). "Other" youth AUC values for individual domains were similar to White, with a 1% greater AUC value for the "G-Factor" (AUC = 0.68). There is less gender distinction, with AUC differences ranging from 0% to 2%. Notably, AUC comparisons are commonly presented when claiming gender and race/ethnicity-neutrality.

To further examine predictive parity, we computed SAR, FPR, and PPV, for gender/sex and race/ethnicity subgroups. For males and females SAR, FPR, and PPV differences were 1% or less. Regarding race/ethnicity, greater variation was observed but when comparing White to minority youth, differences

Table 1. MPACT Model Statistics for MPACT-6 Constructs.

EFA Assessment	*df*	CFI	TLI	RMSEA (90% CI)	SRMR
Education (1-factor)	35	0.966	0.957	0.003 [0.003, 0.003]	0.065
Associations (4-factor)	24	0.969	0.915	0.003 [0.002, 0.003]	0.050
Family (1-factor)	104	0.878	0.859	0.002 [0.002, 0.002]	0.138
Alcohol & Drugs (1-factor)	104	0.967	0.962	0.001 [0.001, 0.002]	0.162
Mental Health (1-factor)	5	0.934	0.868	0.004 [0.003, 0.005]	0.101
Cognition & Behavior (3-factor)	275	0.949	0.944	0.003 [0.003, 0.003]	0.095
CFA Assessment	*df*	CFI	TLI	RMSEA (90% CI)	Omega (ω)
Education	35	0.969	0.961	0.004 [0.004, 0.005]	0.886
Associations	30	0.975	0.962	0.001 [0.000, 0.002]	0.759
Family	101	0.941	0.930	0.002 [0.002, 0.002]	0.871
Alcohol & Drugs	103	0.950	0.942	0.003 [0.002, 0.003]	0.980
Mental Health	1	0.978	0.934	0.007 [0.004, 0.009]	0.892
Cognition & Behavior	272	0.973	0.970	0.002 [0.002, 0.002]	0.968
G-factor	3145	0.934	0.932	0.001 [0.001, 0.001]	0.794

Table 2. MPACT Performance Metrics.

Construct AUCs	Sample	Male	Female	White	Back	Hispanic	Other
Education	0.64	0.65	0.63	0.65	0.62	0.65	0.66
Associations	0.59	0.58	0.60	0.60	0.58	0.56	0.61
Family	0.58	0.58	0.61	0.61	0.57	0.54	0.63
Alcohol & Drugs	0.56	0.56	0.56	0.57	0.54	0.62	0.58
Mental Health	0.52	0.52	0.52	0.52	0.51	0.57	0.51
Cognition & Behavior	0.65	0.65	0.64	0.65	0.63	0.65	0.65
G-Factor Performance							
AUC	0.68	0.68	0.66	0.66	0.66	0.71	0.68
SAR	0.69	0.69	0.69	0.69	0.68	0.67	0.69
False Positive Rate	0.58	0.58	0.59	0.57	0.54	0.62	0.58
Predictive Positive rate	0.57	0.57	0.57	0.57	0.55	0.52	0.57

Note. Due to the large sample size, all models are significant ($p < .001$).

were 5% or less across all metrics. Given the lower relative proportions of Hispanic and “Other” race/ethnic youth, greater emphasis should be given to White and Black youth comparisons. Notably, there were 1% and 2% reductions in SAR and Black youth, indicting slightly reduced performance. However, regarding overclassification, while Black youth had a 2% reduced rate of PPV, this was counterbalanced by a 3% lower FPR, suggesting little-to-no overclassification when comparing these key sub-groups.

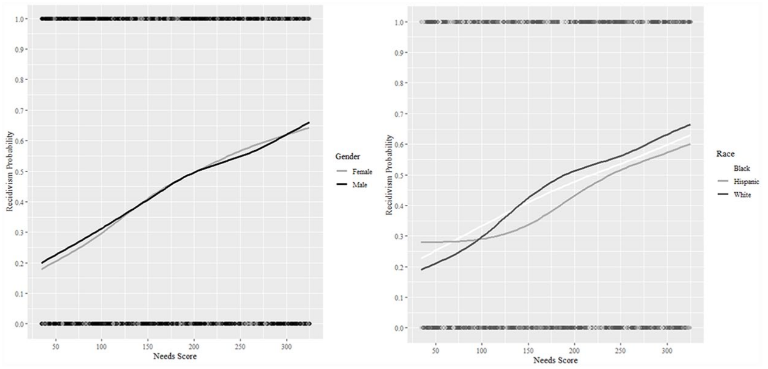

Figure 2. Recidivism and needs score scatter plot by gender and race/ethnicity.

Finally, scatter plots of "G-factor" scores by recidivism probability are presented, with fitted trends for males and females, and White, Black, and Hispanic youth were assessed.[30] Predictive bias is commonly observed in RNA tools, where intercept differences between groups demonstrate lower rates of recidivism for female and minority youth at each point on the risk scale. For gender and race-neutral tools, these "intercept" differences are typically observed throughout the risk scale, demonstrating relatively parallel fit lines that identify over- classification of either females or minority youth (see Hamilton et al., 2019). However, as shown in Figure 2, the recidivism probability for males and females was nearly identical, indicating a .2% over-classification difference (on average) along the needs score continuum. We see a similar trend for youths' race/ethnicity,[31] where the Black fit line (compared to White) indicated a 1% overclassification difference (on average) and did not exceed 5% (~300pts.). When comparing White to Hispanic, the fit line gap did not exceed 8% (~175pts.)32, indicating a 3.5% overclassification difference (on average). Notably, fit lines stayed relatively close, and demonstrating points of trend reversal. Results indicate near prediction parity.

Discussion

Much of the RNA literature has focused on theoretically developed models, pursuing optimal levels of predictive accuracy. To assess risk, RNAs have evolved though multiple generations, combining static risks, dynamic needs, and protective factors in the prediction of an observable outcome—recidivism.

Through RNAs' addition of needs, domains of items were created that, not only worked to improve recidivism prediction, but also provided programing targets to be addressed via correctional interventions. Many agencies attempt to utilize needs scores to determine eligibility criteria, linking domain content to program intent.

However, the assessment of needs is different than risk, where item domains form sub-scales that represent non-observable, latent constructs, require more extensive testing of an assessment's internal structure to assure that needs are "measuring what they intend to measure" (Sullivan et al., 2022). Without these assurances, it is likely that many agencies are inaccurately assessing youth needs and misapplying programing targets. Furthermore, many needs domains were built from the Central Eight and, while a notable theoretical advance at the time, were 1) never fully assessed for construct validity, 2) based on a decades old "brainstorming session," and 3) created for adults, and then applied to JRNAs with little modification or testing. As described, youth often enter the justice system with many needs but few risks and/or a shorter record than adults, where current JRNAs' focus on criminal history, silo needs sub-scores, and have the potential to misclassify those in need of services.

Often practitioners struggle to fully utilize RNAs, where needs domains lack sufficient face validity and cannot be directly tied to established intervention targets (Taxman & Caudy, 2015). Finally, risk assessments, which notably include criminal history indicators, have demonstrated overclassification properties that provide disproportionately higher scores for females and minorities (Angwin et al., 2016; Hamilton et al., 2019). However, many developers have yet to confirm, or neglected to test, these important aspects of needs assessment tools.

The current study sought to assess the MPACT needs assessment. More specifically, following psychometric guidelines, we redesigned the PACT/YASI needs tool. This study takes a critical step, confirming the dimensionality and reliability of the six MPACT needs domains, with a very large sample of youth, across 10-states, and several justice stages (e.g., diversion, probation, parole). Findings demonstrate measurement invariance for both gender/sex and race/ethnicity ensuring the latent needs domains "measure what they are intended to measure." Further, we demonstrate predictive validity of all six domains, across race/ethnicity and gender/sex sub-groups. Based on our review, we believe the MPACT is the first to provide both decisive and robust evidence of construct validity.

Notably, our "G-Factor" provides a much-needed comprehensive assessment of needs to be utilized for programing prioritization (e.g., Taxman & Smith, 2021). Similar to the GAF, our "G-Factor" represents the weighed sum of youth needs across all domains, demonstrating predictive validity strength that rivals that of contemporary risk assessments. This is noteworthy, in that the global needs score was created absent criminal history and other static items, which were thought to hold substantial prediction strength, yet represent an underlying cause of overclassification (Kroner, 2005; W. T. Miller et al., 2022).

Given our positive findings, we believe the MPACT's development moves the RNA needle in several ways and should provide a notable impact for the dozens of agencies currently using the instrument, those seeking to adopt a new instrument, and those using other, contemporary RNAs. First, buttressed by prior findings demonstrating how optimization can improve equity (see Butler et al., 2023; W. T. Miller et al., 2022), the MPACT needs tool supplants the role of bias-inducing criminal history indicators via the expansion and inclusion of dynamic items. While further research should seek to expand our understanding of indicators that contribute to bias, the current study shines a light on the potential importance of dynamic needs and protective items in creating a more equitable prediction for minorities and females.

Second, while the efforts of prior developers provided theoretical justification for latent need domains (Andrews, 1982), differential item weighting (Brennan et al., 2009), and the potential and varying impact of assessment stages (Latessa & Lovins, 2010), a lack of empirical justification for domains and tool design have led to the misclassification of individuals relative to their underlying risk and needs. Potentially this misclassification can result in 1) issues of face validity, neglecting to link domain content to available programing and services; 2) providing services that are detrimental to youth; 3) ineffective recruitment and, in turn, iatrogenic program findings; and 4) subsequently, a lack of practitioner buy-in and drift from assessment best practice applications (Taxman & Caudy, 2015; Viglione et al., 2015). While this list of potential effects of misspecified needs assessments demonstrate a "worst case" scenario, the underlying causes of each may be prevented through proper design and construct validation. Our design of the MPACT needs tool not only outlines the type of reliability and validity evidence agencies should require of their current RNA provider, but we also provide a methodological roadmap to those seeking to redesign existing tools to create new and improved versions.

Third, as youth involved in the justice system continue to decline, it is important to note that reductions in justice system involvement is not likely the result of dramatic declines in the needs of youth. It is more likely that the type of youth previously supervised in the justice system will instead contact social service agencies as a result of their needs (i.e., truancy, substance use, domestic violence, mental health problems and residential instability). This is important, as youth in the justice system have higher needs than the general population, and if left untreated can lead to lifetime disorders such as mental illness, substance abuse, combined with other social determinants of poor health (Elkington et al., 2023). While non-justice interventions may assist in reducing future justice system involvement, social services agencies (e.g., child welfare, dependency) do not commonly provide a similarly diverse and multi-domain assessment of needs, which may result in more youth inadvertently passed over for services needed to improve their quality of life. However, without access to criminal history indicators, JRNA administration is not viable. Notably, the MPACT needs tool provides an opportunity for non-justice agencies to feasibly apply an assessment of youth with potential for future justice system involvement.

Fourth, and importantly, for agencies and researchers to ensure proper development and application, we argue that "best practices" should dictate that needs assessments are created with psychometric criteria. To this point, needs domains are latent and must be developed to include a depth and range of content that are dynamic (changeable over time) and predict recidivism. Second, the domains must be validated to ensure dimensionality, or "measure what they are intended to measure." Third, instruments must demonstrate the needs of youth can be reliably measured, beyond the development sample, including youth from multiple regions and across supervision stages (e.g., diversion, probation, parole). Finally, empirical findings must establish that the content of domains measure needs similarly and provide equitable prediction across gender/sex and race/ethnicity sub-groups.[33] While additional aspects of construct validity (i.e., concurrent, content) may still require assessment, these four aspects are critical for RNAs.

Limitations

Despite attempts to address neglected RNA gaps, the current study is not without limitations. First, we did not create the original item content and were beholden to the responses of the assessment collected. In this vein, feminist scholars have proposed developing and gender-specific assessments for justice-involved girls (Belisle & Salisbury, 2021). While we sought to validate a universally applicable tool for both genders, providing additional items and scales that are gender-responsive may further improve content coverage and prediction.

Second, while there are far more similarities than differences among contemporary JRNA content, there are unique tool elements that may alter the construction of domains. Therefore, the MPACT needs scales may possess assessment content depth and range that is not provided by other tools. With that said, this study represents one of the first to create and empirically validate needs domains, in what we hope researchers, developers, and practitioners will view as a template for similar redesign and development efforts. Further, there are likely differences in terms of how assessments are administered, training received, and interviewing techniques. Unfortunately, we cannot address site fidelity variations in this study. Nevertheless, to mediate this limitation, we equally weighted each state, attempting to diminish the influence of extreme cases, making tool findings more generalizable.

Third, we used a multi-site data set, with some sites having agency of only one justice stage and other sites supervising multiple stages. While we are aware of sites that supervise combinations of diversion, probation, and parole youth, collected measures did not provide identification of youth justice stage. Hence, while we claim broad reliability of measurement findings, we were not able to test measurement invariance and predictive validity across stage. Future research efforts are currently underway to collect and analyze this aspect of the MPACT and we recommend similar assessments of other tools. Also, our findings were restricted to measuring assessment findings at a single time point. We understand that youth needs will change over time and are likely to be influenced by programing and supervision, which was not the focus of the current study. These analyses will be completed in the years to follow, expanding the findings and the described functionality of the MPACT. Nevertheless, upon finishing collecting the data, we will be able to answer this research question in a near future research. In a similar vein, future will test of the tool will explore measurement invariance between youth at different justice stages. The existence of narrow- and multi-band tools attempt to assess risk beyond general recidivism. These tools outline the importance of severity as it pertains to violent and sexual recidivism. Beyond the intent of the current study, additional research is needed to examine the impact of needs tools in the prediction of more specified types of youth outcomes.

Last, from a technical perspective, multitrait-multimethod (MTMM) approach is preferred to establish divergent and convergent validity but is not feasible with the current data structure and sources (D. T. Campbell & Fiske, 1959). While feasibility will always be a difficult hurdle or MTMM, we encourage future researchers to make said attempts using multiple and distinct data collection methods. In addition, we recognized that our statistical models are less-than-perfect. Some of relatively weak item loadings were retained, as they were theoretically important and did not detrimental diminish model fit (see Krushas et al., 2023). Further, we retained these items as we felt it premature remove items that ongoing validation efforts may find these measures substantive and useful as the sample representation continues to expand.

Conclusion

In their seminal work, the *Psychology of Criminal Conduct*, D. A. Andrews and Bonta (2010) outlined overarching principles for effective classification, specifying that a person's risk score, "tells you who to treat" and needs scores, "tells us what to treat" (p.191). While these principles are rooted in modern correctional treatment, we disagree with this described application. Using a risk score to determine a person's level of need is counterproductive, essentially applying a "proxy score" for need that includes static and criminal history indicators that are not impacted by programing and services. This is particularly true for youth, who are younger, may have spent time in another social service system, or do not present extensive offense histories. To resolve this issue, we developed a weighted global score to assess overall need and prioritize programing. We view our "Global Needs Factor" as a sizable achievement with similar utility as other summative latent scales used in other fields (i.e., GAF). In particular, for those agencies with minimal use for a risk assessment, serve cross-over youth, or lack access to criminal history indicators, our "G-factor" may meet their assessment needs yet use fewer resources.

Bonta and Andrews (2016) further state that, "we all have a right to insist upon knowledge. . .that predictions and actions based on (RNAs) are recorded, monitored, and explored empirically" (p.186). We agree. To fully explore neglected validity assumptions, we redesigned an existing needs assessment to provide greater face validity connections with existing programing and services and validated the MPACT's internals structure to give users confidence that the domains "measure what they intend to measure."

Finally, the GPCSL theory, set forth by Bonta and Andrews, indicates that needs domains should measure and predict equally across gender/sex and race/ethnicity (2016). We emphatically agree! With the removal of criminal history and other static risk indicators, the MPACT needs assessment demonstrates measurement invariance and near predictive parity that should ameliorate most concerns of overclassification. With that said, we encourage future RNA examinations to explore beyond underwhelming claims of gender/race-neutrality.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Xiaohan Mei https://orcid.org/0000-0002-7424-8178

Amber Krushas https://orcid.org/0000-0002-4859-7518

Supplemental Material

Supplemental material for this article is available online.

Notes

1. Notably, both the GPCSL and the Central Eight were created in the years that followed the probation officer "brainstorming session," were meta-analytic study reviews were used to justify domain formulation post-hoc (see Wormith & Bonta, 2018).
2. The 10 domains include School, Use of Free Time, Employment, Relationships, Family, Alcohol and Drugs, Mental Health, Attitudes/Behaviors, Aggression, and Skills.
3. Additional generalized RNAs, the Correctional Offender Management Profiling and Alternative Sanctions (COMPAS) and its youth version (Youth COMPAS) were developed with a singular weighted risk score, comprised of mostly justice history items and a separate set of needs domains. Due to the limited availability surrounding the validity of these tools, they are not discussed here.
4. Readers should note that both face and convergent/divergent validity were previously discussed and tested for the MPACT (see *Blinded*).
5. Readers should note that while there are additional tests and aspects of reliability and validity, these are the primary indictors discussed here.
6. Of note, "predictive shrinkage" is a well-known effect, where the strength of an assessments predictive validity is found to reduce substantially when applied to a new sample.
7. These scales include Current School Status, Current Use of Free Time, Current Employment, Current Relationships, Current Living Arrangements, Current Alcohol and Drugs, Current Mental Health, Attitudes/Behaviors, Aggression and Skills.
8. It should be noted that the initial assessment was included for most youth to maximize the number of youths with sufficient follow-up for study inclusion. For confined youth, recidivism exposure in the community is limited and thus, for these youth we utilized their last assessment prior to release from confinement. Further, to retain subject independence, only one assessment per youth was included.

9. Roughly 5.0% of the youths from two of the anonymous states were placed in residents as placement and their last (only) assessment is the study assessment.
10. A more detailed description of all assessment items and responses was presented elsewhere (see *Blinded*).
11. Assessors are trained in accordance with their state's guidelines; however, the content and process of training is not the subject of this study.
12. Missing data procedures were completed via the "misForest" R package, (see *Blinded* for more details).
13. Readers should note that there is about 3.8% of the youths that turn 18 years of age during the follow-up period are still tracked as adult using court administrative records. Also, for the outcome measure, we only used new charges within 12 months, and probation or parole violations were excluded from the analyses.
14. We used EFA approach (Brown, 2015) instead of the preferred approach of multitrait-multimethod matrix (MTMM) (D. T. Campbell & Fiske, 1959) as we were limited to administrative data for the current data structure.
15. All forms of group invariance were tested including configural, metric, scalar, residual, factor, and mean (Schmitt & Kuljanin, 2008).
16. We followed psychometric guidelines for testing sequences of measurement invariance and higher-order factors (see Chen et al., 2005; Rudnev et al., 2017), model specification and identification (see Byrne & Stewart, 2006; Millsap & Yun-Tein, 2004), and omnibus tests (see Little, 2013). Tests were conducted within the Item Factor Analysis (IFA)/Item Response Theory (IRT) framework (Thompson, 2004). As youth were nested within states, a multi-level approach within the structural equation framework (SEM) was used (Matsueda & Drakulich, 2016) using Mplus 8.4 statistical software.
17. The omega was used (instead of Cronbach's Alpha), as it does not assume a parallel construct measurement structure, which is ideal for the current study's data structure (Catalán, 2019; Deng & Chan, 2017). Constructs satisfying (1) dimensionality, (2) measurement invariance, and (3) reliability, were identified to possess structural validity (Rios & Wells, 2014). A threshold of 0.65 for multidimensional (higher order) and 0.80 for unidimensional measures is "acceptable" reliability (Catalán, 2019).
18. Threshold standards include poor (0.32), fair (0.45), good (0.55), very good (0.63) and excellent (0.71) (Tabachnick & Fidell, 2007).
19. EFA findings were used to inform MGCFA models and thus, are not presented in detail. Additional findings may be provided upon request.
20. A Comparative Fit Index (CFI)/Tucker Lewis Index (TLI) of 0.90 or greater and the Root Mean Square Error of Approximation (RMSEA) equal/less than 0.08 is "acceptable"; CFI/TLI are equal/greater than 0.95 and the RMSEA is equal/less than 0.05 is "good" (Brown, 2015; Little, 2013).
21. Readers should note that the chi-square difference test was not employed to compare models, as it is too sensitive to produce reliable results when it is applied to large samples (Cheung & Rensvold, 2002).
22. K-fold advances typical cross-validation, holding out one-tenth of the sample for validation and using nine tenths as a training sample. This process is repeated 10

times, providing an average performance indicated across the 10 validation sets (see Hamilton et al., 2017).

23. To learn about the perceived linkages and relevance to interventions for each domain (see Krushas et al., 2023).
24. Four items were excluded, including “youth needs increasing amounts of alcohol/drug to achieve the same level of intoxication or high” and “youth experiences alcohol/drug withdrawal problems.” These four demonstrated multicollinearity issues (bivariate table of these four items and other items had empty cells); their correlation with others implies these four items are not statistically distinguishable from other items.
25. We did not use the item current suicidal ideation as it did not reach statistically significance thresholds.
26. Standardized item loadings, gender and race/ethnicity invariance test results are reported in a research brief (Krushas et al., 2023).
27. Attendance in most recent term.
28. 1. Person youths live with resulting in an increased risk; 2. Annual combined income of youth & family; 3. Jail/imprisonment history of persons who are currently involved with the household; 4. Problem history of siblings who are currently involved with the household; 5. Youth has run away or been kicked out of home.
29. With large sample sizes, a change that is comparable to and effect size range has been used previously to demonstrate substantive change (see Hamilton et al., 2022).
30. Readers should note that the “Other” race was removed from the plot, as its trend line was inconsistent with model findings. Essentially the low sample size created an unstable fit line that detracted from the findings of White, Black, and Hispanic youth.
31. Of note the proportion of “Other” youth is relatively small (5%) by comparison to White, Black, and Hispanic youth, creating an unstable trend that is not reflective of the underlying findings. While provided in statistical tests, we removed “Other” youth from Figure 2.
32. It should be noted that several states indicated that a less than reliable recoding of Hispanic Ethnicity during several years of data collection, which may have contributed to their lower relative rate.
33. The race/ethnicity of the youth is self-reported.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Andrews. (1982). *A personal, interpersonal and community-reinforcement perspective on devoant behaviour (PIC-R)*. Ontario Ministry of Correctional Services.

Andrews, D. A., & Bonta, J. (1995). *The level of service inventory—Revised*. Multi-Health Systems.

Andrews, D. A., & Bonta, J. (2010). *The psychology of criminal conduct* (5th ed.). Lexis Nexis/Anderson Pub.

Andrews, D. A., Bonta, J., & Wormith, J. S. (2006). The recent past and near future of risk and/or need assessment. *Crime and Delinquency*, *52*(1), 7–27.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. In Martin K. (Ed.), *Ethics of data and analytics* (pp. 254–264). Auerbach Publications.

Baglivio, M. T., & Jackowski, K. (2013). Examining the validity of a juvenile offending risk assessment instrument across gender and race/ethnicity. *Youth Violence and Juvenile Justice*, *11*(1), 26–43.

Barnoski, R. (1997). *Washington State juvenile court recidivism estimates: Fiscal year 1994 youth*. Washington Institute for Public Policy.

Barnoski, R. (2004a). Assessing risk for re-offense: *Validating the Washington State Juvenile Court Assessment (Report No. 04-03-1201)*. Washington State Institute for Public Policy.

Barnoski, R. (2004b). *Washington State Juvenile Court Assessment manual, Version 2.1 (Report No. 04-03-1203)*. Washington State Institute for Public Policy.

Belisle, L. A., & Salisbury, E. J. (2021). Starting with girls and their resilience in mind: Reconsidering risk/needs assessments for system-involved girls. *Criminal Justice and Behavior*, *48*(5), 596–616.

Bonta, J. (2002). Offender risk assessment: Guidelines for selection and use. *Criminal Justice and Behavior*, *29*(4), 355–379.

Bonta, J., & Andrews, D. A. (2016). *The psychology of criminal conduct*. Routledge.

Brennan, T., Dieterich, W., & Ehret, B. (2009). Evaluating the predictive validity of the COMPAS risk and needs assessment system. *Criminal Justice and Behavior*, *36*(1), 21–40.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. New York, NY: Routledge.

Butler, L. C., Hamilton, Z., Krushas, A. E., Kigerl, A., & Kowalski, M. (2023). Racial bias and amelioration strategies for juvenile risk assessment. In E. M. Ahlin, O. Mitchell, & C. Atkin-Plunk (Eds.), *Handbook on inequalities in sentencing and corrections among marginalized populations* (Division on Corrections Sentencing Handbook, Vol. 7, pp. 70–118). New York, NY: Routledge.

Byrne, B. M., & Stewart, S. M. (2006). Teacher's corner: The MACS approach to testing for multigroup invariance of a second-order structure: A walk through the process. *Structural Equation Modeling*, *13*(2), 287–321.

Campbell, C. A., Barnes, A., Papp, J., D'amato, C., Anderson, V. R., & Moses, N. (2020). Understanding the role of neighborhood typology and sociodemographic characteristics on time to recidivism among adjudicated youth. *Criminal Justice and Behavior*, *47*(9), 1079–1096.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105.

Catalán, H. E. N. (2019). Reliability, population classification and weighting in multidimensional poverty measurement: A Monte Carlo Study. *Social Indicators Research*, *142*, 887–910. 2019.

Caudy, M. S., Durso, J. M., & Taxman, F. S. (2013). How well do dynamic needs predict recidivism? Implications for risk assessment and risk reduction. *Criminal Justice Journal*, *41*(6), 458–466.

Chen, F. F., Sousa, K. H., & West, S. G. (2005). Teacher's Corner: Testing measurement invariance of second-order factor models. *Structural Equation Modeling*, *12*, 471–492.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modelling*, *9*(2), 233–255.

Childs, K. K., Peck, J. H., & Brady, C. M. (2022). Predictive bias in juvenile risk assessment: Considering race/ethnicity and sex. *Crime and Delinquency*, 00111287221143936.

Deng, L., & Chan, W. (2017). Testing the difference between reliability coefficients alpha and omega. *Educational and Psychological Measurement*, *77*(2), 185–203. https://doi.org/10.1177/0013164416658325

Elkington, K. S., Robson, G., Sichel, C. E., Lee, J., & Wasserman, G. A. (2023). Family Connect: The Pilot Test of a cross-systems behavioral health treatment referral and linkage intervention for youth on probation. *Criminal Justice and Behavior*, *50*(1), 22–39.

Hamilton, Z., Campagna, M., Tollefsbol, E., van Wormer, J., & Barnoski, R. (2017). A more consistent application of the RNR model: The STRONG-R needs assessment. *Criminal Justice and Behavior*, *44*(2), 261–292.

Hamilton, Z., Kigerl, A., & Kowalski, M. (2022). Prediction is local: The benefits of risk assessment optimization. *Justice Quarterly*, *39*(4), 722–744.

Hamilton, Z., Kowalski, M. A., Kigerl, A., & Routh, D. (2019). Optimizing youth risk assessment performance: Development of the modified positive achievement change tool in Washington State. *Criminal Justice and Behavior*, *46*(8), 1106–1127.

Henning, K. N. (2012). Criminalizing normal adolescent behavior in communities of color: The role of prosecutors in juvenile justice reform. *Cornell Law Review*, 98(2), 383–462.

Herz, D. C., Lee, P., Lutz, L., Stewart, M., Tuell, J., & Wiig, J. (2012). Addressing the needs of multi-system youth: Strengthening the connection between child welfare and juvenile justice. Center for Juvenile Justice Reform and Robert F. *Kennedy Children's Action Corps*. https://cjjr.georgetown.edu/wpcontent/uploads/2015/03/MultiSystemYouth_March2012.pdf

Hollin, C. R., Palmer, E. J., & Clark, D. (2003). The level of service inventory-revised profile of English prisoners: A needs analysis. *Criminal Justice and Behavior*, *30*, 422–440.

Hsu, C. I., Caputi, P., & Byrne, M. K. (2010). Level of Service Inventory–Revised: Assessing the risk and need characteristics of Australian indigenous offenders. *Psychiatry, Psychology and Law*, *17*(3), 355–367.

Jung, S., & Rawana, E. P. (1999). Risk and need assessment of juvenile offenders. *Criminal Justice and Behavior*, *26*(1), 69–89.

Juvenile Justice Geography Policy Practice Statistics. (2020, Nov 23). *The juvenile justice GPS updates risk assessment in probation*. National Center for Juvenile Justice. http://www.jjgps.org/news/article/80/the-juvenile-justice-gps-updates-risk-as.

Kapoor, A., Peterson-Badali, M., & Skilling, T. (2018). Barriers to service provision for justice-involved youth. *Criminal Justice and Behavior*, *45*(12), 1832–1851.

Kitzmiller, M. K., Paruk, J. K., & Cavanagh, C. (2022). *Criminogenic risk score trajectories of Justice-involved youth: An investigation across Race/ethnicity* (p. 00938548221098985). Criminal Justice and Behavior.

Kroner, D. G. (2005). Issues in violent risk assessment: Lessons learned and future directions. *Journal of Interpersonal Violence*, *20*(2), 231–235.

Krushas, A., Hamilton, Z., & Mei, X. (2023). *Assessing the Construct validity of the M-PACT 6 (Document # 02-08-001)*. Nebraska Center for Justice Research. https://www.unomaha.edu/college-of-public-affairs-and-community-service/nebraska-center-for-justice-research/documents/multi-state-confirmation-mpact-2-2023.pdf

Latessa, E. J., & Lovins, B. (2010). The role of offender risk assessment: A policy maker guide. *Victims and Offenders*, *5*(3), 203–219.

Little, T. (2013). *Longitudinal structural equation modeling*. Guilford Press.

Matsueda, R. L., & Drakulich, K. M. (2016). Measuring collective efficacy: A multilevel measurement model for nested data. *Sociological Methods & Research*, *45*(2), 191–230.

Mei, X., Hamilton, Z., Kowalski, M., & Kigerl, A. (2021). Redesigning the Central Eight: Introducing the M-PACT Six. *Youth Violence and Juvenile Justice*, *19*(4), 445–470.

Miller, J., & Maloney, C. (2013). Practitioner compliance with risk/needs assessment tools: A theoretical and empirical assessment. *Criminal Justice and Behavior*, *40*(7), 716–736.

Miller, W. T., Campbell, C. A., Papp, J., & Ruhland, E. (2022). The contribution of static and dynamic factors to recidivism prediction for black and White youth offenders. *International Journal of Offender Therapy and Comparative Criminology*, *66*, 1779–1795.

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, *39*(3), 479–515.

National Academies of Sciences, Engineering, and Medicine. (2022). *The limits of recidivism: Measuring success after prison*.

Olver, M. E., Stockdale, K. C., & Wormith, J. S. (2014). Thirty years of research on the level of service scales: A meta-analytic examination of predictive accuracy and sources of variability. *Psychological Assessment*, *26*(1), 156–176.

Onifade, E., Davidson, W., & Campbell, C. (2009). Risk assessment: The predictive validity of the youth level of service case management inventory with African Americans and girls. *Journal of Ethnicity in Criminal Justice*, *7*(3), 205–221.

Palmer, E. J., & Hollin, C. R. (2007). The level of service inventory— Revised with English women prisoners. *Criminal Justice and Behavior*, *34*(8), 971–984.

Pardoel, K. (2020). *An examination of the influence of gender and race on dynamic risk assessment* [Doctoral dissertation, Carleton University].

Pedersen, G., Hagtvet, K. A., & Karterud, S. (2007). Generalizability studies of the Global Assessment of functioning-split version. *Comprehensive Psychiatry*, *48*(1), 88–94.

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, *41*, 71–90.

Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC area, Cohen's d, and r. *Law and Human Behavior*, *29*(5), 615–620. https://doi.org/10.1007/s10979-005-6832-7

Rios, J., & Wells, C. (2014). Validity evidence based on internal structure. *Psicothema*, *26*(1), 108–116.

Rudnev, M., Lytkina, E., Davidov, E., Schmidt, P., & Zick, A. (2017). Testing measurement invariance for a second-order factor: A cross-national test of the alienation scale. *Methods, data, analyses*, *12*(1), 47–76.

Schmidt, F., Hoge, R. D., & Gomes, L. (2005). Reliability and validity analyses of the youth level of service/case management inventory. *Criminal Justice and Behavior*, *32*(3), 329–344.

Schmidt, N., Lien, E., Vaughan, M., & Huss, M. T. (2017). An examination of individual differences and factor structure on the LS/CMI: Does this popular risk assessment tool measure up? *Deviant Behavior*, *38*(3), 306–317.

Schmidt, S., Heffernan, R., & Ward, T. (2020). Why we cannot explain cross-cultural differences in risk assessment. *Aggression and Violent Behavior*, *50*, 101346.

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review*, *18*(4), 210–222.

Scott, T., Brown, S. L., & Skilling, T. A. (2019). Predictive and convergent validity of the youth assessment and screening instrument in a sample of male and female justice-involved youth. *Criminal Justice and Behavior*, *46*(6), 811–831. https://doi.org/10.1177/0093854819842585

Singh, J. (2013). Predictive validity performance indicators in violence risk assessment: A methodological primer. *Behavioral Sciences & the Law*, *31*.

Singh, J. P., Desmarais, S. L., Sellers, B. G., Hylton, T., Tirotti, M., & Van Dorn, R. A. (2014). From risk assessment to risk management: Matching interventions to adolescent offenders' strengths and vulnerabilities. *Children and Youth Services Review*, *47*, 1–9.

Skeem, J., Monahan, J., & Lowenkamp, C. (2016). Gender, risk assessment, and sanctioning: The cost of treating women like men. *Law and Human Behavior*, *40*(5), 580–593.

Skeem, J. L., Kennealy, P. J., & Hernandez, I. (2013). *CA-YASI construct validity: To what extent do the domains measure the risk factors they're supposed to measure? Division of Juvenile Justice*. California Department of Corrections and Rehabilitation. Rsilience.berkeley.edu/content/publications-1.

Stevenson, H. E., & Wormith, J. S. (1987). *Psychopathy and the level of supervision inventory (User Report 1987-25)*. Ministry of the Solicitor General of Canada.

Steyerberg, E. W., Bleeker, S. E., Moll, H. A., Grobbee, D. E., & Moons, K. G. (2003). Internal and external validation of predictive models: A simulation study of bias and precision in small samples. *Journal of Clinical Epidemiology*, *56*(5), 441–447.

Sullivan, C. J., McCafferty, J., Newsome, J., & Mandalari, A. (2022). Predictive validity and measurement invariance in juvenile risk assessment: Implications for racial and ethnic disparities in juvenile justice. *Journal of Crime and Justice*, *45*, 409–429.

Sullivan, C. J., Strange, C., Sullivan, C., Newsome, J., Lugo, M., Mueller, D., & McCafferty, J. (2019). *Multi-method study on risk assessment implementation and youth outcomes in the juvenile justice system*. University of Cincinnati, Center for Criminal Justice Research.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Allyn & Bacon/Pearson Education.

Taxman, F. S. (2017). *Handbook of recidivism risk/needs assessment tools*. In J. P. Singh, D. Kroner, S. Wormith, S. Desmarais, L. Hamilton, & Z. (Eds.), *Risk assessment: Where do we go from here?* (pp. 269–284). Wiley-Blackwell.

Taxman, F. S., & Caudy, M. S. (2015). Risk tells us who, but not what or how: Empirical assessment of the complexity of criminogenic needs to inform correctional programming. *Criminology & Public Policy*, *14*(1), 71–103.

Taxman, F. S., & Smith, L. (2021). Risk-need-responsivity (RNR) classification models: Still evolving. *Aggression and Violent Behavior*, *59*, 101459. https://doi.org/10.1016/j.avb.2020.101459

Thompson, M. (2004). The value of item response theory in clinical assessment: A review. *Assessment*, *18*(3), 291–307.

Van Voorhis, P., Wright, E. M., Salisbury, E., & Bauman, A. (2010). Women's risk factors and their contributions to existing risk/needs assessment: The current status of a gender-responsive supplement. *Criminal Justice and Behavior*, *37*(3), 261–288.

Viglione, J., Rudes, D. S., & Taxman, F. S. (2015). The myriad of challenges with correctional change: From goals to culture. *European Journal of Probation*, *7*(2), 103–123.

Vincent, G. M., Chapman, J., & Cook, N. E. (2011). Risk-needs assessment in juvenile justice: Predictive validity of the SAVRY, racial differences, and the contribution of needs factors. *Criminal Justice and Behavior*, *38*(1), 42–62.

Vincent, G. M., Guy, L. S., & Grisso, T. (2012). *Assessment in juvenile justice: A guidebook for implementation*. John D. and Catherine T. MacArthur Foundation.

Wong, S. C. P., & Gordon, A. (2006). The validity and reliability of the Violence Risk Scale: A treatment-friendly violence risk assessment tool. *Psychology Public Policy and Law*, *12*(3), 279–309.

Wormith, S. J., & Bonta, J. (2018). *The level of service (LS) instruments. Handbook of recidivism risk/needs assessment tools* (pp. 117–145). John Wiley & Sons.

Zottola, S. A., Desmarais, S. L., Lowder, E. M., & Duhart Clarke, S. E. (2022). Evaluating fairness of algorithmic risk assessment instruments: The problem with forcing dichotomies. *Criminal Justice and Behavior*, *49*(3), 389–410.

Author Biographies

Xiaohan Mei is an Assistant Professor of Criminal Justice and Criminalistics at California State – Los Angeles. His research focuses on risk and needs assessment for correctional populations.

Zachary Hamilton is an Associate Professor of Criminology and Criminal Justice and the Associate Director of the Nebraska Center for Justice Research at the University of Nebraska – Omaha. His research focuses on risk and needs assessment for correctional populations.

Alex Kigerl is a Senior Research Associate for the Nebraska Center for Justice Research at the University of Nebraska – Omaha. His research focuses on risk and needs assessment for correctional populations.

Amber Krushas is an Assistant Professor of Criminal Justice at the University of Nevada – Las Vegas. Her research interests include juvenile justice and victimology.

Faye S. Taxman is a University Professor, Director of the Center for Advancing Correctional Excellence! at George Mason University. Her research interests include implementation science, program fidelity, and assessment application.