Youth risk assessment approaches:
Lessons learned and question raised by Baird et al.'s study (2013)

Comment by Jennifer Skeem and (in alphabetical order) Robert Barnoski, Edward Latessa, David Robinson & Claus Tjaden*

*This is the **final** version of this comment. NCCD refused to print this version – instead, in Baird et al. (2013) they printed a draft provided before the target report was complete.*

For information, contact:
Jennifer L. Skeem, Professor
School of Social Welfare; Goldman School of Public Policy
University of California, Berkeley
jenskeem@gmail.com

OVERVIEW

Context and purpose

In juvenile justice agencies across the U.S., it has become common to apply structured tools to assess a youth's risk of re-offending and/or to inform efforts to reduce that risk. For good or for ill, an industry has grown up around "risk-needs" assessment and states increasingly are developing their own "risk assessments." Many risk assessment tools are now available. Although most tools stem from the same root, they vary in their degree of complexity, structure, and independent research support. These tools, in turn, are being implemented in agencies that differ in their levels of organizational commitment to both the value(s) of risk assessment, and the necessity of ensuring that staff has adequate training, skills, and motivation to score the tools correctly.

Given this diversity of tools and implementation efforts, the time is ripe for a snapshot of the reliability and utility of risk assessment in juvenile justice agencies. That snapshot has just been provided for several agencies, in the form of a study by Baird et al. (2013).

We are delighted that Baird et al. (2013) conducted this study. We believe that their data provide a valuable picture that can be used to advance "real world" risk assessment. We are concerned, however, that their presentation of these data will promote mistaken conclusions. The field should not abandon an entire, relatively new approach to risk assessment because some tools have some problems in some jurisdictions -- that would amount to throwing the baby out with the bathwater.

Before beginning, it is important to note who we are. This comment was written by four of the five advisory board members who participated in the final meetings held in Baltimore, where Baird et al.'s (2013) report was discussed at length (i.e., all attending advisory board members except Howell). The fifth coauthor (Latessa) could not attend those meetings. Like Baird (who helped create the Solano County instruments, the GSC and Girls Link), three of us have a conflict of interest because we are directly attached to a tool/approach evaluated in this study. Some of these tools performed well, as implemented in this study….others did not. The primary author of this rebuttal (Skeem) and the final coauthor (Latessa) are professors with no such conflict of interest.

This comment focuses on "big picture" issues most relevant to policy-makers and practitioners. We leave aside specific methodological problems with Baird et al.'s (2013) report that may have affected the results.[1]

---

[1] For example, the CRN was developed with one scoring method for adjudicated and probated youth, but the authors disaggregate the two samples; the YASI has a prescreen, but the authors develop their new scale using items from the full instrument; the PACT combines two subscales into a single risk assessment, but the authors present AUCs for two subscales as if they are independent (where a single AUC for the sum of subscales better represents the PACT).

Summary of key points

In this commentary, we articulate four conclusions that can be drawn from this study. We then present the fundamental question that this study cannot address. These key points are as follow:

- **Conclusion 1: There is room for improvement in both risk assessment tools AND the quality with which they are implemented.** Although Baird et al. (2013) tend to attribute their findings solely to tools, their study cannot disaggregate the quality of a tool from the quality with which it was implemented. At the broadest level, their results indicate that a variety of tools, *as implemented in a variety of sites*, have room for improvement in their reliability and predictive utility.

- **Conclusion 2: Inter-scorer reliability is not self-evident.** In almost half of the sites studied, staff was unable to score the tool in a manner that was consistent with that of an expert. When staff scores a tool incorrectly, the tool's ability to inform accurate decisions about youth is limited. Inter-rater reliability cannot be ignored during processes of development or implementation.

- **Conclusion 3: Risk classifications must be cross-validated and/or customized.** Above all, this study provides a compelling reminder that agencies must check and "customize" risk classifications (e.g., low, medium, high), based on local sample characteristics. Based on differences in youth populations and recidivism rates, one agency's high-risk case may be another agency's low-moderate risk case. When classifications are not fit to an agency, the predictive utility of an otherwise accurate tool will be forsaken in everyday practice.

- **Conclusion 4: Short tools can predict as well as (not better than) longer ones.** Most of Baird et al.'s (2013) report seems allocated to the argument that "shorter is better" and that the "Solano GSC is best." The data do not support these conclusions. The tools with the greatest predictive utility, as implemented in this study, were the Oregon JCP (31 items), YASI Virginia (32 items), and Solano GSC & Girls Link (9 items). Like past studies, this study indicates that short tools sometimes predict as well as longer ones. Similar levels of predictive utility can be achieved by (a) statistically selecting and combining a few highly predictive risk factors, and (b) sampling risk domains more broadly and including risk factors that can inform risk reduction efforts.

- **Open question: What value is added by risk-reduction oriented approaches?** Contemporary risk assessment approaches are oriented toward the *prediction of recidivism,* the *reduction of recidivism,* or both. Tools oriented solely toward prediction tend to be simpler than those oriented toward reduction. Baird et al.'s (2013) study raises a question that it cannot address: What evidence is there that reduction-oriented risk assessment tools add value to those that are prediction-oriented? For reduction-oriented tools, it is not enough merely to demonstrate that adding variables "does no harm" to predictive utility. Precious juvenile justice resources should not be spent on pointless assessment exercises.

Instead, these tools must demonstrate that the variables they add actually bring something of value to the risk reduction enterprise. There are several potential avenues for doing so. It is time for the field to get serious about addressing this important and challenging question.

CONCLUSIONS SUPPORTED BY DATA

There is room for improvement in risk assessment tools and/or their implementation

At the broadest level, the results of this study indicate that a variety of risk assessment tools, *as implemented* in a variety of sites, have room for improvement in their reliability and predictive utility.

Baird et al.'s (2013) opinion aside, the "Area Under the ROC Curve" or AUC is the most appropriate statistic for comparing the predictive utility of tools across sites. In part, this is because -- unlike the DIFR -- its size is not affected by base rates of recidivism, which range from 11% to 51% across sites in this study.

Only one tool at one site – Oregon's JCP – achieved an AUC of .70, the minimum level of predictive accuracy "considered acceptable for clinical application purposes" (Zhang, Roberts, & Farabee, in press, p. 5). As shown in Baird et al.'s (2013) Table 40, five tools/sites manifested a "medium" effect in predicting readjudication (i.e., AUC $\geq$ .649), four manifested a "small" effect (i.e., AUC $\geq$ .556), and four essentially had no effect. None of the tools/sites achieved a large effect size (AUC $\geq$ .712).[2] A similar picture emerges with Baird's et al.'s (2013) preferred statistic, the DIFR.

This study cannot pull apart the quality of a risk assessment tool from the quality with which it is implemented. Although Baird et al. (2013) tend to attribute their findings solely to instruments, each finding also reflects implementation quality.[3] To identify high quality tools for the field (on one hand), and guidelines for implementing them (on the other), future work should attempt to differentiate between these two issues. This would allow researchers and practitioners to develop guidelines for (a) demonstrating that a tool is well-validated before it is disseminated, and (b) adequately implementing well-validated risk assessment tools.

---

[2] As shown by Rice and Harris' analyses (1995), minimum AUCs of .556, .639, and .712 correspond to "small," "medium," and "large" effect sizes, respectively.

[3] For example, based on results for two YLS/CMI sites included in the present study, Baird et al. (2013, p. 51) conclude that "the YLS/CMI appears to have limited value as a classification tool". Nevertheless, a large body of peer-reviewed research provides more favorable results for the predictive utility of the YLSI/CMI. The discrepancy between Baird et al.'s (2013) findings and past research are consistent with the well-validated correctional principle that implementation quality matters.

Inter-scorer reliability is *not* self-evident

This study examines a critical, but routinely ignored issue: inter-scorer reliability. When staff score a risk assessment tool in an inconsistent or incorrect manner, that tool cannot inform accurate decisions about youth. Reliability is a necessary (but not sufficient) condition for a tool to accurately predict recidivism. It is, therefore, a key element of evidence-based practice in risk assessment.

Baird et al. (2013) found that staff provided with exactly the same information about a youth were able to attain "good" scoring agreement with other staff in 9 of 10 study sites,[4] but attained adequate scoring agreement with an expert in only 5 of the 11 study sites.[5] In other words, staffs' scores are often consistent with one another, but not necessarily "correct."[6]

Reliability problems typically reflect poorly defined items and/or inadequately trained staff. Both causes seem to be culprits, here. First, across tools (from the GSC Solano to the YASI Virginia), items that were abstract and/or poorly defined tended to be less reliable. This suggests that tool developers must define items carefully and empirically demonstrate that they can be scored reliably. Second, staff at different sites scored the same tool with different levels of reliability.[7] This suggests that the quality of training and implementation matters – in keeping with a large body of correctional treatment research. Agencies should train their staff until they attain a specified level of reliability, and then periodically reassess whether staff are scoring the tool correctly.[8]

Risk classifications must be cross-validated and/or customized

Above all else, the results of this study provide a compelling reminder that agencies must check and "customize" risk classifications, based on local sample characteristics (see Andrews &

---

[4] See Table 38, Column 6. "Good" is defined as an ICC > .75, following guidelines by Parkerson, Broadhead & Tse (1993). (Because it is not appropriate to compute ICCs for ordinal data, the ICCs reported for "risk levels" in Table 38, Column 5 are questionable.)

[5] See Table 38, Column 4, which depicts the average proportion of staff scores that exactly match expert scores across items. "Inadequate" is defined as < 75%.

[6] This possibility could be tested by using consensus scores generated by an expert panel of scorers as the criterion for staff, rather than scores provided by a single individual.

[7] See Table 38, where the YLS/CMI attains an ICC of .80 in Nebraska, but only .67 in Arkansas.

[8] These two factors probably interact. Even though research indicates that they robustly predict criminal behavior, abstract risk factors like criminal attitudes or poor parental supervision are harder to measure than concrete risk factors like criminal history. Tools probably vary in how well they measure those abstract risk factors. Sites vary in how well they train and monitor staff. When an abstract risk factor manifests poor predictive utility on a tool within a site, is that a fault of the tool, a problem with its implementation in that site, or both? Without additional information, it will be impossible to tell.

Bonta, 2003). Risk classifications involve nothing more – and nothing less – than chopping up a continuous score on a risk assessment tool to create a number of ordinal categories (e.g., "low," "medium," "high"). Tool developers often use a particular sample of youth to optimize risk classifications– i.e., identify cut scores that create reasonably sized groups of youth with recidivism rates that are as different as possible. Using the language of the DIFR statistic, one goal is to maximize "base rate dispersion."[9]

The problem is that risk classifications that are optimized in one sample can degrade when they are applied to a new sample – particularly when the new sample has a much different risk score distribution, base rate of recidivism, or both. Based on differences in their youth populations and recidivism rates, one agency's high-risk case may be another's average bear.

This underscores the necessity of locally assessing and validating the predictive utility of risk assessment scores and classifications. In some cases, risk classifications will not be meaningful unless they are customized. One sign that this is the case is when the predictive utility of scores (as indexed by the AUC) outstrips the discrimination ability of classifications (as indexed by the DIFR). Based on the sites studied by Baird et al. (2013; see Table 40), this "outstripping" happens often enough to be concerned. Specifically, risk assessment scores moderately predicted new adjudications in five sites (i.e., AUC $\geq$ .639).[10] Although risk classifications also performed well in three of these five sites (i.e., high DIFR for Oregon JCP, GSC Solano, YASI Virginia), they performed poorly in the remaining two (i.e., low DIFR for Girls Link Solano, CRN Georgia).[11]

For example, in Georgia, there is a 64% probability that a (randomly selected) adjudicated youth will obtain a higher CRN score than a (randomly selected) non-adjudicated youth (AUC=.64). So, CRN scores do a moderately good job of distinguishing between youth with- and without- a new adjudication. However, CRN classifications performed relatively poorly (DIFR= .40). Specifically, there wasn't much difference between "moderate" and "high" groups in their adjudication rates. This is a sign that the agency needs to customize cut scores to their sample. If the agency uses risk classifications that do not fit their sample to inform decision-making about youth, then they are forsaking the predictive utility of scores on that tool. The "high" risk youth isn't meaningfully different from the "moderate" risk youth.

Assuming that *scores* on the tool are predictive in the new agency, the good news is that risk classifications can be modified to fit the new agency's population. Ideally, an agency would modify risk classifications not only to maximize their base rate dispersion, but also to fit the decision(s) that they want classifications to inform. For example, if the goal is to identify low

---

[9] Silver, Smith, & Banks (2000).

[10] See note 1 above for AUC interpretation guidelines.

[11] No interpretation guidelines (e.g., "small," "medium," "large") are available for the DIFR. Users must be cautious in applying the DIFR because its size is affected by recidivism rates.

risk cases to divert from detention, then (a) only two risk classifications are needed ("low" and "not low"), and (b) the cut score can be adjusted (within limits) to be lower or higher, to reflect that agency's weighting of public safety, youth rights, and resource concerns.

In short, this finding is an important call to the field to get serious about cross-validating and (if necessary) customizing risk classifications to their setting. As Baird et al. (2013) note, agencies tend to use classifications more than scores to inform their decision-making about youth. These results suggest that researchers and policy-makers should articulate guidelines for cross-validating and customizing risk classifications. Ensuring that risk classifications are valid is essential, when implementing any risk assessment tool.

Short tools can predict recidivism as well as (not better than) longer ones

Some of the tools included in this study are relatively short and simple (i.e., the GSC Solano, 9 items and Arizona AOC, 9 items); most others are relatively long and/or complex (like the YASI Virginia, 32 items). Loosely, these tools represent an evolution in risk assessment over time, from prediction-oriented approaches (which were designed solely to achieve efficient prediction) to reduction-oriented approaches (which also emphasize variable risk factors that theoretically can be changed to reduce risk).[12]

Most of Baird et al.'s (2013) report seems allocated to the argument that "shorter is better" and that the "GSC Solano is best."[13] In their introduction, the authors caution, "If changes to risk assessment instruments have resulted in diminished capacity to accurately discriminate among high-, moderate-, and low-risk-youth, then decision-making in juvenile justice has been adversely affected." In addition to planned analyses that test the reliability and predictive utility of each instrument at each site, the authors perform extensive post hoc analyses in an attempt to (a) create shorter and (ideally) more predictive versions of relatively long tools, and (b) to create a GSC proxy that (ideally) predicts better than rival tools. The authors conclude that "the GSC, used in Solano County, proved to be the most successful risk instrument evaluated in this study;" that their GSC Solano proxy "transferred better" than rival tools; that "complex scoring systems…diminish results," and that most shorter instruments they created "produced *markedly better* results than the instrument currently in use."

The study's results do not support these conclusions. First, although the results of planned analyses indicate that predictive utility varies across sites (see AUCs and DIFRs in Table 40), there is no evidence that this variability is a simple function of a tools' length or complexity.[14]

---

[12] For a review of this evolution and the confusion it has created, see Monahan & Skeem (in press).

[13] This argument is apparent in the authors' review of past research, which is highly selective. For example, the authors select only the least favorable finding (from 20 largely positive comparisons), when referencing results from a New York YASI sample (Orbis, 2007).

[14] Baird et al. (2013) could directly test the relationship between tool length (and/or complexity) and predictive utility by performing a small meta-analysis with their data. We did not do so because they do

For example, scores on both the YASI Virginia and GSC Solano – this study's prototypes of "long/complex" and "short/simple" -- manifested good inter-rater reliability (ICCs = .89-.92; see Table 38) and equivalent predictive utility (AUC and DIFR=. 68 for both; see Table 40). Indeed, the tools with the most predictive scores and classifications were a locally created tool (the Oregon JCP, 31 items), a simple public domain tool (the GSC Solano, 9 items), and a "later generation" commercial tool (the YASI Virginia, 32 items).

Second, at best, the results of the authors' post hoc analyses demonstrate that short tools can predict re-adjudication as well as longer ones. The authors created new tools for ten sites to maximize prediction within each dataset by (a) using statistical criteria to select and combine variables, and then (b) customizing risk classifications (see above). [15] As a rule, tools constructed in this way capitalize on chance associations between variables in a particular sample and will "shrink" in predictive power when applied to new samples. So, tools must be cross-validated with an independent sample. [i]

Baird et al. cross-validated three of the ten tools they created. The authors did not test whether their new tools predicted re-adjudication significantly better than the original tools (by testing differences in AUCs -- or any other statistic). In fact, for some instruments, they did not even provide estimates of predictive utility that could be directly compared (e.g., inconsistently separating estimates by gender). Nevertheless, as shown in Table 1 below, there generally is little difference in the predictive utility of the original ("longer") tools and cross-validated new ("shorter") tools. The average AUC difference is .02. The average DIFR difference was also a modest .13 – and this difference may be based more on customization of risk classifications than any substantive change to the tool (see Footnote 15). The only direct comparison that can be made is for the Arizona AOC, where performance is essentially equivalent. Incidentally, the new scale had *more* items than the original Arizona AOC scale.

---

not operationalize either variable in their report (i.e., item number, item/scoring complexity)…and there are too few exemplars of short/simple tools (i.e., two) to support an adequate test.

[15] **Unfortunately, the authors conflated the development of new tools with the customization of classifications**. Their analyses would have been much more informative if they had customized risk classifications based on original scores to assess the degree of improvement this yielded, before developing new scores and risk classifications that were tightly fitted to a particular dataset. (This is another reason the AUC is a more comparable indicator of performance across tools than the DIFR – particularly in this study.)

Table 1. Similar Predictive Utility for Original Tools and Baird et al.'s (2013) New Tools

Assessment	Number of items, original vs. new tool(s)	Original tool's utility (boys & girls combined)	New tool's utility, cross-validation sample
Arizona AOC Risk Asst. Inst.	9 vs. 12	AUC= .62 DIFR = .40	AUC= .63 DIFR= .42
PACT Florida Probation	22 vs. 12 (boys) & 11 (girls)	AUC= .59 "Criminal"; .63 "Social" DIFR = .37	AUC= .66 boys only; .66 girls only DIFR= .53 boys only; .52 girls only
CRN Georgia	59 vs. 9	AUC= .64 DIFR= .40	AUC= .67 boys only DIFR = .61 boys only

A similar case is apparent when the performance of original instruments is compared with the GSC Solano proxies. As shown in the table below, the "longer" tools perform about as well as the "shorter" tools. It is highly unlikely that AUC differences of .02 or .03 are statistically significant. Even the DIFR differences were minimal, at .07 and .01 (and these reflect local customization of risk classifications as much as any substantive tool change).

Table 2. Similar Predictive Utility for Original Tools and GSC Solano Proxy

Assessment	Number of items, original vs. new tool	Original tool's utility	GSC Solano Proxy utility
PACT Florida Probation	22 vs. 12	AUC= .60 "Criminal"; .62 "Social" DIFR = .37	AUC = .63 DIFR= .44
CRN Georgia	59 vs. 9	AUC = .64 DIFR= .47	AUC = .66 DIFR= .48

Fundamentally, this study provides evidence that tools that differ in their length, format, and foci can achieve similar levels of predictive utility. This finding is consistent with research on the relative predictive utility of alternative risk assessment tools that, as a group, are much better validated than those studied here.[16] Despite heated debate about which type of tool predicts best ('actuarial' vs. 'clinical;' simple vs. complex; etc.), research is making it increasingly clear that there is no winner in this horse race. For example, in a meta-analysis of 28 separate studies, Yang, Wong, and Coid (2010) found that the predictive efficiencies of nine validated risk assessment instruments were essentially ''interchangeable,'' with estimates of accuracy falling within a narrow band (AUC= .65 to .71). The tools examined included a short actuarial device that emphasizes simple risk markers (the VRAG), a more clinically oriented tool that

[16] For a meta-analysis, see Olver, Stockdale, & Wormith (2009).

emphasizes variable risk factors (the HCR-20) -- and virtually everything in-between (like the LSI-R).

Two factors may help explain the similar predictive performance of well-validated instruments. First, these tools seem to tap ''common factors'' or shared dimensions of risk, despite their varied items and formats.[17] Second, these tools seem to reach a "glass ceiling" of predictive utility, beyond which they cannot improve. If a limiting process makes recidivism impossible to predict beyond a certain level of accuracy, each tool can reach that limit quickly with a few maximally predictive items, before reaching a sharp point of diminishing returns. Baird et al.'s (2013) post hoc results are consistent with this possibility, and echo the results of other studies. For example, based on a sample of over 1,000 released prisoners, Coid et al. (2011) found that most individual items included in risk assessment tools did not significantly predict violence. When these items were removed, the resulting reduced scales predict violence as well as (but usually not better than) the original full scale. For example, a 5-item version of a prediction-oriented scale (the VRAG) performed as well as the full 12-item version (AUCs= .70, .71, respectively). It is important to recognize that if there is a glass ceiling, it can be reached via alternative routes. If measured validly, some variable risk factors (e.g., attitudes supportive of crime) predict recidivism as strongly as common risk markers (e.g., early or "pre-adult" antisocial behavior; Gendreau et al., 1996).

In short, similar levels of predictive utility can be achieved by (a) statistically selecting and combining a few highly predictive risk factors, and (b) sampling risk domains more broadly and including risk factors that can inform risk reduction efforts. For these reasons, Skeem and Monahan (2011) concluded:

> "Given a pool of instruments that are **well validated** for the groups to which an individual belongs, our view is that the choice among them should be driven by the ultimate purpose of the evaluation. If the ultimate purpose is to characterize an individual's likelihood of future [criminal behavior] relative to other people, then choose the most efficient instrument available. This is appropriate for a single event decision in which there is no real opportunity to modify the risk estimate based on future behavior. If the ultimate purpose is to manage or reduce an individual's risk, then value may be added by choosing an instrument that includes treatment-relevant risk factors... This choice is appropriate for ongoing decisions in which the risk estimate can be modified to reflect ebbs and flows in an individual's risk over time."

[17] In an innovative demonstration, Kroner, Mills, and Reddon (2005) printed the items of four well-validated instruments (e.g., LSI-R, VRAG) on strips of paper, placed the strips in a coffee can, shook the can, and then randomly selected items to create four new tools. The authors found that the ''coffee can instruments'' predicted violent and nonviolent offenses as well as the original instruments did. Factor analyses suggested that the instruments tap four overlapping dimensions: criminal history, an irresponsible lifestyle, psychopathy and criminal attitudes, and substance-abuse-related problems. Each of these dimensions were predictive of recidivism.

OPEN QUESTION: DOES REDUCTION-ORIENTED RISK ASSESSMENT ADD VALUE?

At its core, the study by Baird et al. (2013) raises a fundamental question that it cannot address: What evidence is there that reduction-oriented risk assessment tools add value to those that are prediction-oriented? It is time for the field to get serious about addressing this important and challenging question.

At the risk of oversimplification, Baird et al. (2013) mistakenly assume that the only purpose of risk assessment is classification; and the only real measure of a tool's performance in meeting that purpose is predictive utility (i.e., base rate dispersion). Their yardstick of success is defined by parsimony and predictive utility. Period.[18]

This yardstick is both sensible and sufficient, when the ultimate purpose of risk assessment is merely to characterize a youth's likelihood of recidivism, compared to other youth. In this case, what the tool assesses is irrelevant because there is no interest in explaining or reducing risk. For example, if a tool that that efficiently assesses accuracy in playing street dice strongly predicts recidivism (see Nunnally, 1978), then the tool is valid for characterizing risk. As summarized by Gottfredson and Moriarty (2006), "if a variable can be measured reliably, and if it is predictive, then of course it should be used—absent legal or ethical challenge."

When the ultimate purpose of risk assessment is to reduce a youth's risk of recidivism, predictive utility is a necessary -- but not sufficient -- measure of success. Contemporary thinking and "later generation" risk assessment tools have been infused with the concepts of risk management and risk reduction. Theoretically, these tools add value to simple tools by assessing variable risk factors[19] (e.g., antisocial attitudes; poor parental supervision) that may help explain the process that leads to recidivism. The goal is to inform risk reduction efforts by

---

[18] The evolution of correctional risk assessment tools has created a largely artificial distinction between "risk" and "needs" assessment (see Monahan & Skeem, in press). "Risk" assessment tends to be reduced to an actuarial formula that heavily weighs risk markers. Sometimes the items that comprise this formula are explicitly separated from other items (e.g., Baird's JAIS/GSC), and sometimes they are embedded among other items (e.g., YASI, PACT, CRN). "Needs" assessment tends to be whatever content remains on the tool, once the predictive items have been removed. Baird et al.'s (2013) evaluation criteria imply that the "risk" part of these tools is subject to scientific scrutiny, but "anything goes" for needs assessment. If the field follows this suggestion, few gains will be made in understanding and reducing risk among youth. Instead, we believe that the field can and should evaluate whether these tools – in their entirety -- are capable of fulfilling their intended purposes.

[19] Variable risk factors are variables that have been shown to predict recidivism and to be changeable (see Monahan & Skeem, in press for clear definitions of fixed markers, variable markers, variable risk factors, and causal risk factors). Sometimes variable risk factors are called "dynamic risk factors" or "criminogenic needs."

(a) specifying risk factors to target in treatment, and (b) capturing any changes in risk over time to inform ongoing decisions about supervision and treatment.

Baird et al.'s (2013) yardstick is not sufficient for measuring the success of these tools. All risk assessment tools must manifest adequate predictive utility…but this only gets "later generation" tools to first base. For these tools, it is not enough merely to demonstrate that adding variables "does no harm" to predictive utility. Precious juvenile justice resources should not be spent on pointless assessment exercises. Instead, these tools must demonstrate that the variables they add actually bring something of value to the risk reduction enterprise. There are several potential avenues for doing so. For example, one could test a tool's construct validity to determine whether it actually measures the variable risk factors that it says it measures (for an example, see Skeem, Kennealy & Hernandez, 2013). Or test whether variable risk factors assessed by a tool change over time, and whether those changes predict recidivism (for an example, see Dixon & Howard, 2013). Or test in a well-controlled study whether youth are significantly less likely to recidivate when professionals use a reduction-oriented rather than prediction-oriented assessment approach. The most rigorous (and treatment-relevant) test would be a randomized controlled trial in which a targeted intervention was shown to be effective in changing a variable risk factor(s) on a tool, and the resulting changes were shown to reduce the likelihood of post-treatment recidivism (see Monahan & Skeem, in press).

Practice has far outpaced research, at this intersection between risk assessment and risk reduction. An absence of evidence that these tools add value to risk reduction efforts, however, is not the same as counter evidence. We strongly recommend that researchers and policy-makers work together to articulate concrete measures for testing the value added by reduction-oriented risk assessment tools. The time could not be better to take on this challenge, given the current level of interest in using science to inform real problem-solving in the juvenile justice system.

References

Andrews, D. A., & Bonta, J. (2003). *The psychology of criminal conduct* (3rd ed.). Cincinnati, OH: Anderson.

Baird, C., Healy, T., Johnson, K., Bogie, A., Dankert, E., & Scharenbroch, C. (2013). *A Comparison of Risk Assessment Instruments in Juvenile Justice*. Madison, WI: National Council on Crime and Delinquency.

Coid, J. W., Yang, M., Ullrich, S., Zhang, T., Sizmur, S., Farrington, D., & Rogers, R. (2011). Most items in structured risk assessment instruments do not predict violence. *The Journal of Forensic Psychiatry & Psychology, 22(1*), 3-21.

Gendreau, P., Little, T., & Goggin, C. (1996). A meta-analysis of the predictors of adult offender recidivism: What works. *Criminology, 34(4),* 575–608.

Gottfredson, S. D., & Moriarty, L. J. (2006). Statistical risk assessment: Old problems and new applications. *Crime and Delinquency, 52*(1), 178–200.

Howard, P., & Dixon, L. (2013). Identifying change in the likelihood of violent recidivism: Causal dynamic risk factors in the OASys violence predictor*. Law & Human Behavior,* 37, 163-174.

Kroner, D. G., Mills, J. F., & Reddon, J. R. (2005). A coffee can, factor analysis, and prediction of antisocial behavior: The structure of criminal risk. *International Journal of Law and Psychiatry, 28*(4), 360–374.

Monahan, J., & Skeem, J. (in press). Risk redux: The resurgence of risk assessment in criminal sentencing. *Federal Sentencing Reporter.*

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Olver, M. E., Stockdale, K. C., and Wormith, J. S. (2009). Risk assessment with young offenders A meta-analysis of three assessment measures. *Criminal Justice and Behavior, 36(4),* 329-353.

Orbis Partners. (2007). *Long-term validation of the Youth Assessment and Screening Instrument (YASI) in New York State Juvenile Probation.* Ottawa, Ontario: http://criminaljustice.state.ny.us/opca/pdfs/nyltyasifullreport20feb08.pdf

Parkerson, G. R., Eugene Broadhead, W., & Tse, C.-K., J. (1993). The Duke Severity of Illness Checklist (DUSOI) for measurement of severity and comorbidity. *Journal of Clinical Epidemiology, 46*(4), 379–393.

Rice, M. E., & Harris, G. T. (1995) Violent recidivism: Assessing predictive validity. Journal of Consulting and Clinical Psychology, 53, 737–748.

Silver, E., Smith, W. R., & Banks, S. (2000). Constructing Actuarial Devices for Predicting Recidivism A Comparison of Methods. *Criminal Justice and Behavior*, *27*(6), 733-764.

Skeem, J. L., & Monahan, J. (2011). Current directions in violence risk assessment. *Current Directions in Psychological Science, 20(1),* 38-42.

Skeem, J., Kennealy, P., & Hernandez, I. (2013). *CA-YASI construct validity: To what extent do the domains measure the risk factors they're supposed to measure?* Available at: http://riskreduction.soceco.uci.edu/index.php/publications-all/published-risk-assessments/

Yang, M., Wong, S. C., & Coid, J. (2010). The efficacy of violence prediction: A meta-analytic comparison of nine risk assessment tools. *Psychological Bulletin, 136(5),* 740-767.

Zhang, S., Roberts, R., & Farabee, D. (in press). An Analysis of Prisoner Reentry and Parole Risk Using COMPAS and Traditional Criminal History Measures. *Crime & Delinquency.* Published online at http://cad.sagepub.com/content/early/2011/11/16/0011128711426544.full.pdf+html

Endnote

[i] Table X. compares the predictive utility of the original tools and of Baird et al.'s new tools that were not cross-validated. Estimates for the new tools are likely inflated because the same sample was used to optimize and "test" the new tool. Still, the pattern of results suggests that tools with moderate predictive utility were difficult to improve, regardless of their length. The un-validated new tools generally did not predict recidivism better than the original Oregon JCP, GSC Solano, Girls Link Solano, or YASI Virginia (Average AUC difference = .02). There was more room for improvement among scales with weaker utility (Arizona DJC and YLS/CMI Nebraska, Average AUC difference = .09). The degree of improvement appears unrelated to the degree of shortening.

Table X. Predictive utility of original tools and non-cross-validated new tools

Assessment	Number of items, original vs. new	Original tool AUC	New tool, construction AUC	AUC difference
Arizona DJC DRI	18 vs. 15	.59	.69	.10
YLS/CMI Nebraska Probation	42 vs. 16	.55	.61	.06
YLS/CMI Nebraska Commit.	42 vs. 11	.54	.66	.12
Oregon JCP	31 vs. 12	.70	.70	.00
GSC Solano County	10 vs. 9	.68	.70	.02
Girls Link Solano County	10 vs. 9	.68	.73	.05
YASI Virginia	32 vs. 15 (boys) & 11 (girls)	.68	.71 (boys) & .74 (girls)	-- non-nested