ACLU Colorado

October 29, 2020

Dangerous, Misleading and Biased: A Letter on Pretrial Risk Assessment Tools in Colorado

Dear Colleague,

We have closely reviewed the July 1, 2020 Colorado Pretrial Assessment Tool Validation Study published by the University of Northern Colorado and funded by the State of Colorado (hereinafter "UNC Study").[1] We write now to highlight some of the underreported and concerning findings of this study, and sound the alarm regarding the new risk assessment instrument proposed in the study, referred to as the CPAT-R. We are gravely concerned that the CPAT-R takes Colorado — once considered a leading pretrial reform state — in the wrong direction.

As detailed further below, the UNC Study reveals:

1. **Discrimination in Colorado's pretrial risk assessment tools**
    - Colorado's current pretrial risk assessment tool (CPAT) unfairly discriminates against Black people and people experiencing homelessness by more often mistakenly identifying Black people and homeless people as "high risk," when compared to White people and housed people.
    - Data from the UNC Study suggests that CPAT-R may be even more discriminatory against Black people and people experiencing homelessness than the CPAT.
    - This discriminatory effect is particularly unfair in light of UNC data showing that Black people and White people engage in pretrial misconduct at equal rates.

2. **Poor predictive capacity of both the CPAT and CPAT-R**
    - The CPAT validated at the lowest acceptable level for a risk assessment tool and performs marginally at predicting pretrial "failure".
    - While the CPAT-R is better than the CPAT at correctly assigning people to the low-risk categories, it performs *worse* than the CPAT at accurately predicting who belongs in the highest risk category.
    - Mistaken placement of individuals in high risk categories can have dire consequences, including pretrial detention and onerous conditions of release.

3. **Failure to predict behaviors that matter in pretrial decision making**
    - The CPAT could not be validated to predict likelihood of violent behavior or flight from prosecution, which should be the key consideration in pretrial decision making.
        - **The pretrial population is not a danger to the public**. Part of the reason the CPAT could not be validated to predict violence is because violence is rare among the pretrial population. Less than 2% of people released on bond are charged with a violent offense during the pretrial period.
    - Both the CPAT and CPAT-R could only be validated to predict marginally relevant pretrial behavior — whether a released defendant may miss a single court appearance or may be charged with an arrestable offense (the vast majority of which are low-level misdemeanors). Thus, these tools are of limited, if any, utility.

4. **The CPAT-R wildly overestimates people's risk**
    - CPAT-R's highest risk category is populated by low risk people. In fact, 74% of the "highest risk" people are likely to avoid any criminal charge if released pretrial, and 66% are likely to appear at every single court appearance.
    - A tool that identifies as high risk so many individuals who succeed pretrial is, quite simply, useless.

Taken together, these shortcomings raise fundamental questions about the usefulness and fairness of continuing to use the CPAT in Colorado. The CPAT-R, however, is not a viable alternative. In ways that matter, it is less accurate and potentially more discriminatory than the CPAT, while at the same time overestimating the risk of people released pretrial. It is our belief that adopting the CPAT-R could very well increase the incarcerated pretrial population.

I. Introduction - *Overview of Pretrial Risk Assessment Tools & the UNC Study*

A. What is a pretrial risk assessment tool and how do they work?

A pretrial risk assessment tool is an automated decision system that purports to predict the likelihood of a person's future actions if released on bond, such as whether they will appear for court hearings or be charged with an arrestable offense. Risk assessment tools output a risk score for each individual, which are then turned into categories in decision matrices that provide guidance for pretrial decision-makers. Often, those in higher categories, supposedly representing higher risk of pretrial "failure," are recommended for pretrial detention or the most onerous pretrial conditions, with less restrictive conditions as the categories lower. In recent years, these tools gained prominence throughout the country as an ostensibly data-driven, unbiased mechanism to assess a defendant's pretrial risk and facilitate safely lowering the incarcerated

pretrial population. However, an increasing amount of evidence suggests serious flaws in these tools both in their design and implementation, including:

1. **Discriminatory impact**, baked into the tools themselves, against people of color and other groups overrepresented in the criminal legal system, including people living in poverty, and people with disabilities;[2]

2. Problems with how decision makers interact with the tools — notably judges provided risk assessments scores have shown **increased bias in their pretrial decision making**;[3] and

3. Demonstrable **failure to reduce incarceration rates** in many instances.[4]

With these concerns in mind, some of the once staunchest defenders of risk assessment tools are now acknowledging the tools as flawed and harmful beyond repair.[5]

Nonetheless, most Colorado jurisdictions use a pretrial risk assessment tool. The most commonly used Colorado tool is the Colorado Pretrial Assessment Tool (CPAT), and there is a push for all Colorado counties to adopt and use this tool. The CPAT assigns a risk score from 1 to 4, and judges, prosecutors, public defenders and pretrial service staff across the state rely on that score to urge for and make decisions regarding bond conditions, often including the amount of monetary bond a defendant must pay to secure pretrial freedom. People deemed higher risk by the CPAT (categories 3 and 4) are more likely to have a high monetary bond set or onerous conditions of release set, such as GPS monitoring. Because monetary bonds frequently result in individuals without financial means remaining incarcerated, the CPAT score often influences whether or not someone will remain in jail pretrial. In other words, **the CPAT has a direct impact on people's liberty and their ability to defend themselves against criminal charges. In that sense, the stakes could not be higher. If Colorado is to use a pretrial risk assessment tool, it had better get it right.**

B. The UNC Study confirms fatal flaws in the CPAT and CPAT-R.

Unfortunately, as the UNC Study revealed, the CPAT is a tragically flawed instrument. The study showed the CPAT is only marginally competent at predicting outcomes. Worse, as advocates have long urged, the data confirms that the CPAT unfairly discriminates against Black people and people experiencing homelessness. Perhaps realizing that Colorado cannot continue to use a tool that is both marginal at prediction and unfairly discriminatory, the UNC Study validates and suggests broad usage of a new tool called CPAT-R. Unfortunately, the CPAT-R creates more problems than it fixes. While both the CPAT and CPAT-R suffer from critical flaws, including near random predictions of *genuine* public safety outcomes and bias against protected classes, we believe the CPAT-R poses a substantial risk of *increasing* the number of people jailed pretrial. Moreover, the CPAT-R is worse than the CPAT at accurately predicting

who belongs in the highest risk categories, with grave consequences for those who are erroneously deemed high risk.

Below we detail several critical concerns regarding both the CPAT and the CPAT-R, and we urge extreme caution in moving forward with using the revised CPAT-R in any Colorado jurisdiction.

II. Analysis

A. The UNC Study confirms that both the CPAT and CPAT-R are discriminatory.

As the Pretrial Justice Institute (PJI) recently explained its shift from heralding pretrial risk assessment tools to wholesale rejecting them: "Regardless of their science, brand, or age, these tools are derived from data reflecting structural racism and institutional inequity that impact our court and law enforcement policies and practices. Use of that data then deepens the inequity."[6] The ACLU stands with PJI, racial justice organizations, community advocates, formerly incarcerated people, and many data analysts and scholars in rejecting the use of algorithmic risk assessment tools in pretrial decision making.[7] The tools are racist at their foundation. That is because the data underlying these tools are historical criminal justice data. "Including criminal history in the tool might seem reasonable, but doing so ignores the fact that racial biases, not necessarily behavior, often determine whether someone gets a criminal record."[8] Indeed, "arrests, charges, bail amounts, and sentencing are all more harshly meted out against Black people, when compared with White people."[9]

There is a statistics idiom for this kind of flawed data and its impact on risk assessments: "Garbage in, garbage out." This means that an "algorithmic prediction is only as good as the data on which the algorithm is trained."[10] With an input of racist policing data into the algorithm, one should expect a racist output.[11] Specifically, one should expect that these tools will unfairly assign higher risk scores to overpoliced Black and Brown communities, who resultantly face a greater likelihood of pretrial incarceration. As discussed below, that is precisely what we know is happening for the CPAT, and there is strong evidence that such discriminatory effects are present for the CPAT-R as well.

Discriminatory risk assessment tools are often more dangerous than using no tool at all. This is because such tools, although biased in their outcomes, are presented as neutral, objective measures of pretrial success. **In that sense, these tools, "although they may seem objective or neutral – threaten to further intensify unwarranted discrepancies in the justice system and to provide a misleading and undeserved imprimatur of impartiality for an institution that desperately needs fundamental change."[12]**

The UNC Study confirms what advocates have long urged: the CPAT unfairly discriminates against Black people. Racial justice advocates have long insisted that risk

assessments disproportionately overestimate the risk of failing in the pre-trial period for Black people, and the CPAT study confirms this.

As an initial matter, the UNC Study confirms that Black people released pretrial are no more risky than White people. Indeed, the rates of pretrial "failure" for Black people and White people in the study sample were *nearly identical,* with 30-31% of Black and White people failing pretrial, regardless of race.[13] Nonetheless, the CPAT places Black people in the higher risk categories more often than White people, which results in a higher "false positive" rate for Black people than White people.[14] What this means is that the CPAT overpredicts pretrial failure by Black people more often than for White people. This error occurs 1.2 times, or 20% more often for Black people than for White people.[15] **Taken together, equal base rates and disparate false positive rates imply that Black people will suffer bias when this tool is used, because they are more likely to be wrongly classified as high risk and thus more likely to wrongly be subject to pretrial detention or onerous conditions of release.**

The CPAT discriminates against people experiencing homelessness. To its credit, the UNC Study is one of the few analyses to address the question of differences in errors for people experiencing homelessness. The UNC Study concludes, as advocates have long urged, that the CPAT is unfairly biased against people experiencing homelessness. Specifically, the tool dramatically overestimates the risk of pretrial failure for homeless people, with nearly 79% of unhoused people in the highest risk category succeeding pretrial compared to a 51% success rate of those who are housed.[16] Errors in predictions of pretrial failure for people experiencing homelessness occurs 1.5X, or 50% more often, than for housed individuals.

While homelessness is not itself a protected class, homelessness is a close proxy for many disability status types.[17] Additionally, because people experiencing homelessness most often do not have the means to pay even the lowest monetary bond, a risk assessment tool that is biased against homeless people *and* is used to inform the decision on whether and how high to set money bond, raises serious constitutional concerns. The Equal Protection and Due Process Clauses prohibit incarceration for poverty.[18] Yet, a risk assessment tool that unfairly overestimates the risk level of people experiencing homelessness will predictably result in higher bond amounts and — therefore — more frequent incarceration for this group. At minimum, excessive assessments of risk for unhoused people implies that indigent defendants continue to face heightened scrutiny and barriers to release because of their poverty. Notably, the UNC Study does not analyze the intersection of race and homelessness, but we suspect being Black and homeless will only increase the odds of being erroneously identified as high-risk by the tool.

The CPAT-R maintains the same biases against Black and homeless people as the CPAT. Perhaps recognizing that the CPAT's biases render the tool too flawed for use, the UNC authors piloted, validated and recommended a new tool, called "CPAT-R." While the authors suggest that their new tool is bias neutral, their conclusion is based on analysis of a limited set of test

data. The validation sample set reflects even more bias against Black people and people experiencing homelessness than the CPAT. As **Figure 1** shows, Black people in the top 2 categories (considered the highest risk categories) are incorrectly predicted to fail 1.5X times, or 50% more often than White people in the same category, and homeless people are incorrectly predicted nearly 2X more often than housed people.[19] The harms of discriminatory false positives in the two riskiest categories are substantial — people who score higher on the tool are more likely to be detained pretrial or, if released, to have liberty-restricting conditions of release. **When it comes to bias against Black people and people experiencing homelessness, the UNC Study buries the lead: the CPAT-R is no better than CPAT — both are discriminatory**.

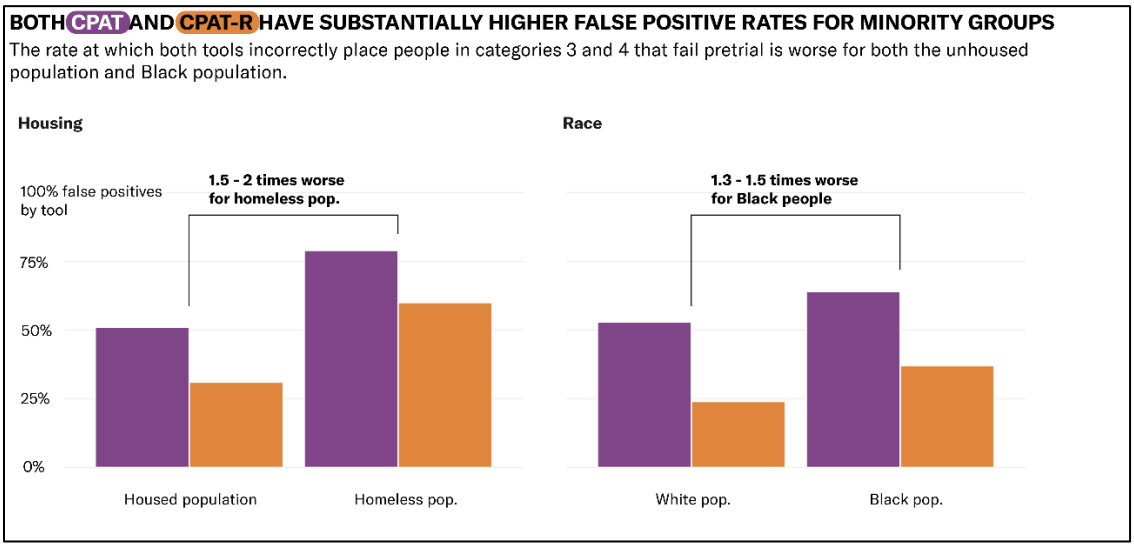

Figure 1. *Chart showing CPAT and CPAT-R false positive rates for minority groups.*

The CPAT-R likely introduces additional bias problems with its over-emphasis of failure to appear in court (FTA). As shown in **Figure 2**, most of the factors considered by CPAT-R substantially overpredict the risks presented by Black and Latinx people, when compared to White people. However, CPAT-R's heavy emphasis on past FTAs is likely the biggest driver of the discriminatory false positives. This emphasis is concerning on its face given admissions throughout the UNC Study that many FTAs are non-willful or due to no fault of the defendant.[20] Despite admissions on all sides that non-willful FTAs are common, and limited evidence they are relevant to public safety or flight risks, the CPAT heavily penalizes anyone who has had a single FTA in the last year.[21] The bias data in the UNC Study reflects that this unnecessary and heavily weighted factor substantially contributes to overprediction of risk for both Black and Latinx people (see Figure 2).[22] Importantly, this particular bias problem is new with the CPAT-R, because the CPAT did not consider past FTAs.

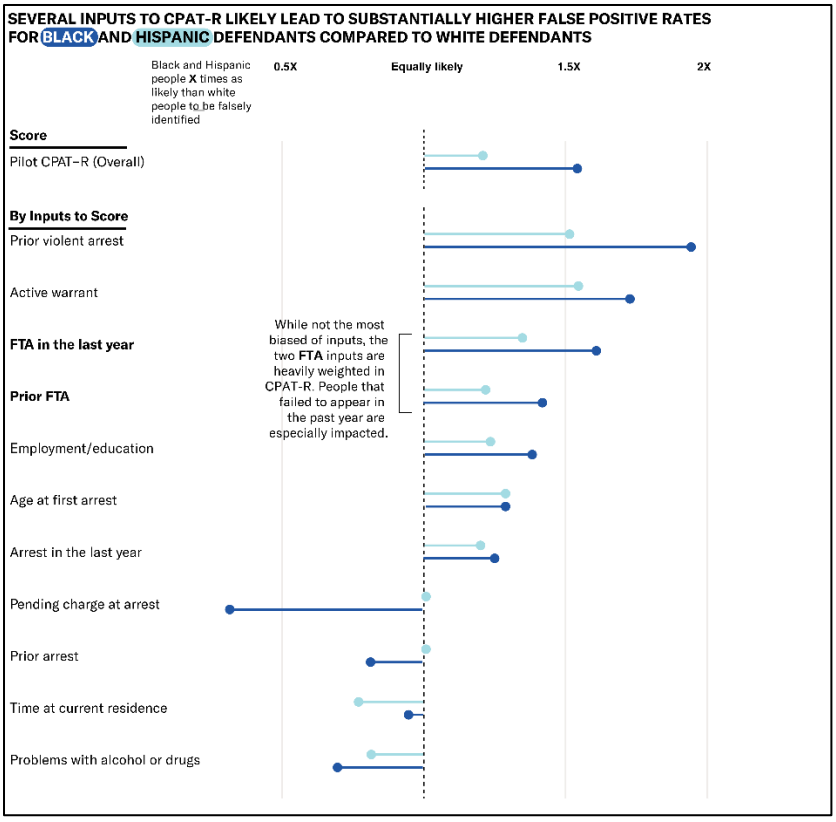

Figure 2. *Chart showing CPAT-R inputs that lead to higher false positive rates for Black and Latinx defendants.*

B. The CPAT and CPAT-R fail to predict what the courts should care about — violence and flight from prosecution — but the authors gloss over this most essential failure in both tools.

The CPAT has long been the subject of criticism because it attempts to predict behavior that should have little impact on pretrial decision making. Specifically, it attempts to predict only whether a pretrial defendant, if released: (1) will miss a **single court appearance** (whether willful or not); or (2) **will be charged with an arrestable offense**. Yet, courts have long admonished that, when it comes to pretrial liberty, the considerations that matter are avoiding: (1) flight from prosecution or willful failure to appear; and (2) serious and violent criminal offenses.[23] The CPAT has never attempted to meet this standard.[24] As the UNC Study recognizes, a single missed court appearance most certainly does not constitute flight from prosecution, or even willful failure to appear. Similarly, many arrestable offenses are neither serious nor violent. Indeed, data from Colorado's Division of Criminal Justice confirms that more than 50% of arrestable offenses charged while people are on bond are for (1) traffic misdemeanors, (2) drug possession, and (3) misdemeanor assault.[25] The risk that a defendant, if released, might commit one of these offenses or miss a single court appearance cannot constitutionally justify pretrial detention. **Putting the Constitution aside, it is morally**

abhorrent to deprive a presumptively innocent person of their pretrial liberty based on a concern they might get a traffic ticket or fail to show up in court one day. Yet, high CPAT scores — *designed* to measure these very risks — often are a causal factor in pretrial detention. That is because, in practice, judges who see a high risk score often set money bonds in a misguided attempt to "offset" the risk.

Perhaps in an attempt to finally address this criticism, the UNC authors studied whether the CPAT could be used to predict the pretrial risks that matter: risk of violence and risk of willful failure to appear. The UNC Study reflects that the CPAT is unable to accurately predict violent offenses. The reason for this is simple. As the authors candidly explained in a recent Pretrial Executives Network Meeting, the pretrial population is an extremely low risk population as a whole. Indeed, the rates at which pretrial people commit violent offenses are so low that there is no risk assessment tool in the country that does well at predicting these outcomes.[26] The UNC Study reflects that less than 2% of pretrial defendants in the sample dataset were charged with a violent offense while out on bond.[27] **With so few people charged with a violent crime while on bond, the UNC Study found that the CPAT was no better at predicting violence than chance.**[28]

When it comes to missed court appearances, however, the study authors could have strived to make a tool that directionally points to the relevant outcomes. They did not do so. In the UNC Study, the authors did attempt to differentiate FTAs to get at the concept of willfulness. They broke FTAs into three categories: (1) no consequence FTAs, where the court takes no action in response to the FTA; (2) low consequence FTAs, where the court imposes a moderate sanction; and (3) high consequence FTAs, where the FTA is followed by a form court sanction, such as warrant issuance.[29] The base rate of "high consequence" failures — those that result in warrants by the court — is 20.23%, a base rate high enough to focus on in the risk assessment validation. [30] The authors chose not to do so, and instead kept FTAs (which include non-willful FTAs) as both their primary outcome variable and as an input into the risk assessment itself.

Figure 3 is a graphical representation of the predictive performance of the CPAT and CPAT-R related to subcategories of FTAs (e.g. high consequence vs. low consequence) and arrestable offenses (e.g. violent or serious). The only individual variable for which CPAT-R outperforms CPAT with high confidence is the FTA High Consequence variable — where the confidence intervals (lines on the plot) for CPAT fall outside of the range of the CPAT-R. For domestic violence re-arrest and violent re-arrest, both CPAT and CPAT-R are essentially random as their confidence intervals overlap a 0% Gini index (the dotted line, representing completely at random guesses).[31] For all other re-arrest variables, predictions are barely better than random guessing (distance from the top of the plot, which represents a perfectly accurate prediction instrument).

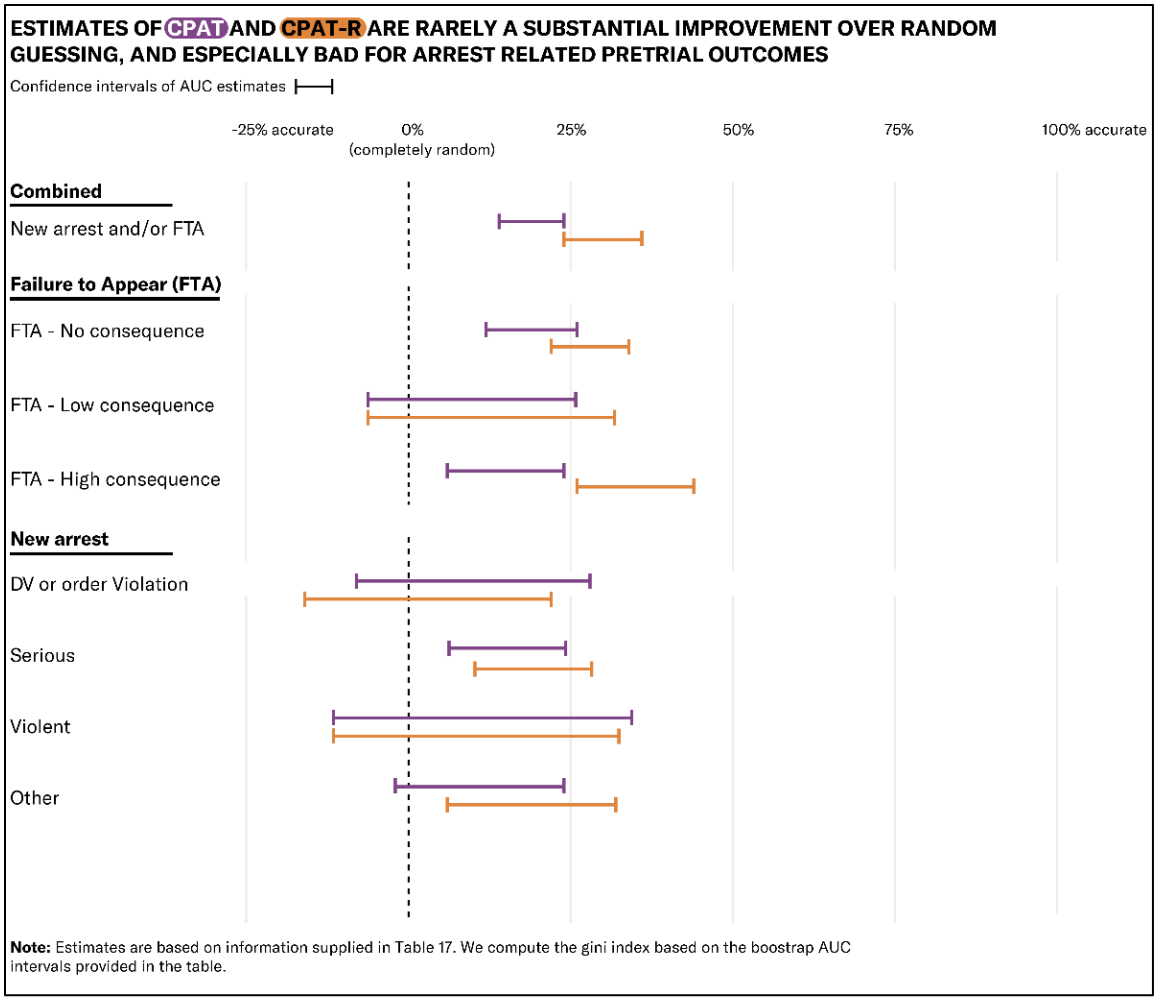

Figure 3. *Chart showing CPAT and CPAT-R's ability to accurately predict failure to appear or a new arrest compared to random chance.*

In the end, instead of grappling with the valid and relevant pretrial risks — violence and flight — the new CPAT-R sticks to the flawed risks measured by the CPAT — a single missed court appearance or being charged with *any* arrestable offense.[32] These risks are marginally relevant in the pretrial context and should never warrant restrictions on liberty. Taken together with the now proven biases in the tool, **it is time to question, why does Colorado continue to cling to pretrial risk assessment tools?**

C. Neither the CPAT nor the CPAT-R are good at predicting pretrial "failure."

The UNC Study reveals that neither the CPAT nor the CPAT-R are good at predicting pretrial failure, even if broadly defined as a single missed court appearance or being charged with an arrestable offense. The UNC Study validates the CPAT as adequately predictive and offers the CPAT-R as having better predictivity than the CPAT. In doing so, the authors rely almost exclusively on Area Under the Curve (AUC) both to assess tool accuracy and justify the choice

of the CPAT-R over the CPAT.[33] The UNC Study reflects that the CPAT is only marginal at predicting outcomes. With a .58 AUC, **the CPAT validated near the bottom of the lowest acceptable performance range ("fair") that can justify validation.**[34] The authors urge that the CPAT-R has improved accuracy, with an AUC score above .60[35]. A closer look at the data, however, reveals that, **while the CPAT-R is somewhat better at predicting low risk people than the CPAT, the CPAT-R is substantially worse at predicting high risk people**. As noted above, the consequences of a high risk score can be severe, including incarceration, GPS monitoring and more. Thus, failures at predication in the higher risk categories can be catastrophic for individuals.

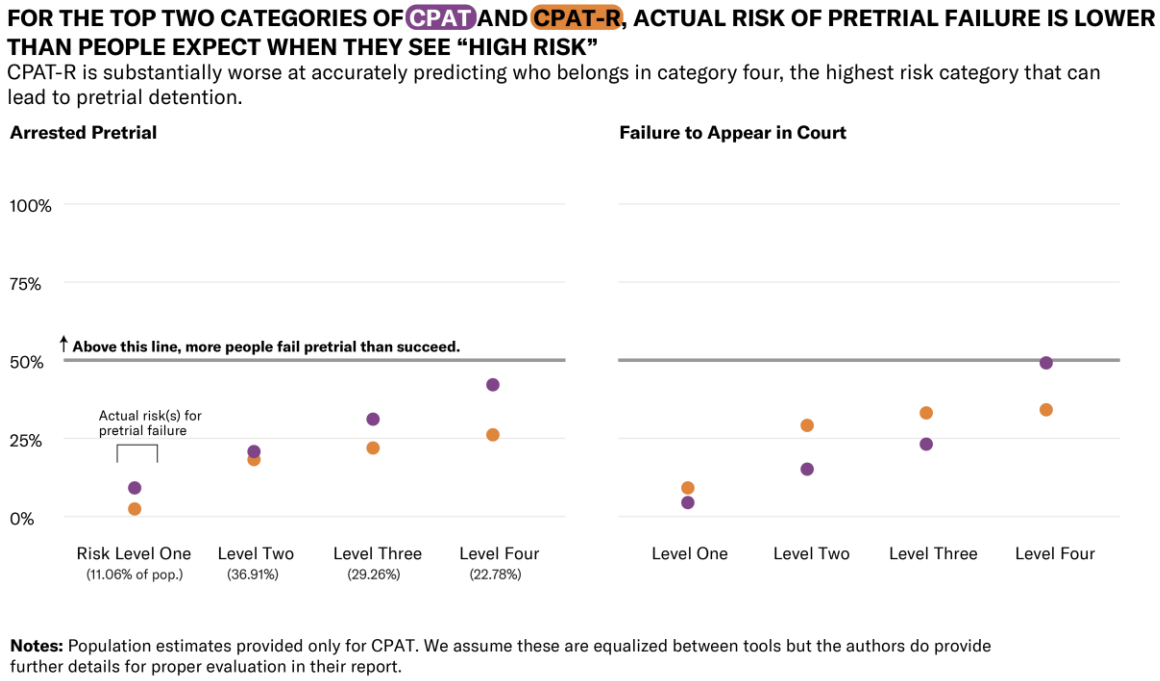

Figure 4. *Chart showing that actual risk of pretrial failure is lower than expected for both the CPAT and CPAT-R*

D. The CPAT-R's "higher risk" categories are wildly misleading.

Perhaps of greatest potential impact on the growth of the pretrial population, the CPAT-R's higher risk categories (categories 3 and 4) are wildly misleading. That is because CPAT-R's high risk categories are populated by a surprisingly large percent of people who *succeed* pretrial. As **Figure 5** shows below, in Category 4, 74% of people released are predicted to commit no new arrestable offense, including even a traffic misdemeanor. Likewise, 66% of Category 4 individuals are predicted to appear at every single court date without fail. Such success rates do not meet any arguable definition of "high risk." Yet, it is likely that pretrial decision makers will continue to use high risk designations by the CPAT-R as justification for onerous conditions of release and/or higher monetary bond. As a result, presumptively innocent people who are at

relatively low risk of pretrial "failure," will likely be unnecessarily subject to pretrial detention or liberty restricting conditions of release due to the CPAT-R.

Importantly, the CPAT-R's overly inclusive high risk categories are substantially worse than those in the CPAT. As **Figure 5** shows, for each outcome — court appearance and new arrestable offense — CPAT-R people in the highest risk category (category 4) are significantly more likely to succeed pretrial than compared to the CPAT.[36] In that sense, the CPAT-R is substantially more misleading on the issue of risk than the CPAT, much to the detriment of individuals erroneously deemed high risk by the tool.

Risk Category	CPAT New Arrestable Offense, Rate of Success	CPAT-R New Arrestable Offense, Rate of Success	CPAT Court Appearance, Rate of Success	CPAT-R Court Appearance, Rate of Success
1	91%	97%	95%	91%
2	80%	82%	85%	71%
3	69%	78%	77%	67%
4	58%	74%	51%	66%

Figure 5. *CPAT-Rs highest risk category (category 4) includes individuals who are much less "risky" than CPAT. Rates of success in the highest risk category (74% no new arrestable offense; 66% appear for every court date) do not comport with any common understanding of a "high risk" individual.*

Given that the higher risk categories of CPAT-R include many more individuals who are successful pretrial than compared to the CPAT, we expect widespread adoption of the CPAT-R poses a serious risk of increasing Colorado's pretrial population, and unnecessarily so.

To the extent proponents of the CPAT-R argue that will not happen — that judges will understand, based on the data, that the Category 4 people are not actually high risk, that begs the questions: "What is the purpose of a high risk category where the majority of people are not re-arrested pretrial and where most appear just fine for their court dates? [37] How does knowing that an individual has a high chance of succeeding pretrial at the highest risk category serve to meaningfully inform pretrial decision making? The answer: it does not.

III. Conclusion

Pretrial detention and onerous conditions of release are often driven – in substantial part—by risk assessment tools. Yet, the evidence is clear that Colorado's CPAT, and even the proposed

CPAT-R are biased and misleading **tools that serve as a crutch to replace thoughtful, individualized judicial decision making.**

We urge Colorado judges, prosecutors, public defenders and pretrial service providers to move away from pretrial risk assessment tools altogether and to wholly avoid adoption of the CPAT-R, which risks substantially increasing our detained pretrial population.

We invite further conversation and dialogue on this important topic.

Sincerely,

Rebecca T. Wallace
Senior Staff Attorney and Senior Policy Counsel
ACLU of Colorado

Aaron Horowitz
Deputy Chief Analytics Officer
ACLU National

[1] Victoria A. Terranova and Kyle C. Ward, *Colorado Pretrial Assessment Tool Validation Study Final Report*, UNIVERSITY OF NORTHERN COLORADO (Jul. 1, 2020), https://www.nacdl.org/getattachment/18510570-e0eb-4d40-b737-5aafb30c1085/terranovaward_cpat-validation-study_final-report.pdf

[2] Matt Henry, *Risk Assessments: Explained*, THE APPEAL (Mar. 25, 2019), https://theappeal.org/risk-assessment-explained; Laurel Eckhouse et al., *Layers of Bias*, SAGE PUBLICATIONS (Nov. 23, 2018), https://journals.sagepub.com/doi/10.1177/0093854818811379; Sandra G. Mayson, *Bias In, Bias Out*, YALE LAW JOURNAL (Sept. 28, 2018), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3257004

[3] Brian P. Schaefer and Tom Hughes, *Examining Judicial Pretrial Release Decisions: The Influence of Risk Assessments and Race*, CRIMINAL JUSTICE, LAW & SOCIETY (Aug. 1, 2019), https://ccjls.scholasticahq.com/article/9908-examining-judicial-pretrial-release-decisions-the-influence-of-risk-assessments-and-race; Alex Albright, *If you Give a Judge a Risk Score: Evidence from Kentucky Bail Decisions*, HARVARD UNIVERSITY (Sept. 3, 2019), https://thelittledataset.com/about_files/albright_judge_score.pdf; Ethan Corey, *How a Tool to Help Judges May Be Leading them Astray*, THE APPEAL (Aug. 8, 2019), https://theappeal.org/how-a-tool-to-help-judges-may-be-leading-them-astray/; Megan T. Stevenson and Jennifer L. Doleac, *Algorithmic Risk Assessment Tools in the Hands of Humans*, SOCIAL SCIENCE RESEARCH NETWORK (Dec. 5, 2019), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3489440.

[4] See earlier references; also see Megan T. Stevenson and Jennifer L. Doleac, *The Roadblock to Reform*, AMERICAN CONSTITUTION SOCIETY (Nov. 2018), https://www.acslaw.org/wp-content/uploads/2018/11/RoadblockToReformReport.pdf; Megan T. Stevenson, *Assessing Risk Assessment in Action*, SOCIAL SCIENCE RESEARCH NETWORK (Aug. 27, 2017), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3016088; Ethan Corey, *New Data Suggests Risk Assessment Tools Have Little Impact On Pretrial Incarceration*, THE APPEAL (Feb. 7, 2020), https://theappeal.org/new-data-suggests-risk-assessment-tools-have-little-impact-on-pretrial-incarceration/.

[5] *See, e.g.*, *Updated Position on Risk Assessment Tools*, PRETRIAL JUSTICE INSTITUTE (Feb. 7, 2020), https://www.pretrial.org/wp-content/uploads/Risk-Statement-PJI-2020.pdf; Ben Green, *The False Promise of Risk Assessments: Epistemic Reform and the Limits of Fairness*, HARVARD UNIVERSITY (Jan. 2020), *https://scholar.harvard.edu/files/bgreen/files/20-fat-risk.pdf;* Bryce Covert, *A Bail Reform Tool Intended To Curb Mass Incarceration Has Only Replicated Biases in the Criminal Justice System*, THE INTERCEPT (Jul. 12, 2020), *https://theintercept.com/2020/07/12/risk-assessment-tools-bail-reform/*

[6] *Updated Position on Risk Assessment Tools*, PRETRIAL JUSTICE INSTITUTE (Feb. 7, 2020), https://www.pretrial.org/wp-content/uploads/Risk-Statement-PJI-2020.pdf

[7] *The Use of Pretrial "Risk Assessment" Instruments: A Shared Statement of Civil Rights Concerns*, PRETRIAL JUSTICE INSTITUTE (Jul. 30, 2018), http://civilrightsdocs.info/pdf/criminal-justice/Pretrial-Risk-Assessment-Full.pdf

[8] *Racist Risk Assessment Algorithms Should Not Be the Future of Sentencing in Pennsylvania*, THE INQUIRER (Sept. 4, 2019), https://www.inquirer.com/opinion/editorials/risk-assessment-algorithm-tool-pennsylvania-sentencing-commission-20190904.html; Laurel Eckhouse et al., *Layers of Bias*, SAGE PUBLICATIONS (Nov. 23, 2018), https://journals.sagepub.com/doi/10.1177/0093854818811379, see p. 196-97, which discusses problems in assessing bias of a risk assessment given the many biases inherent in the data it relies on.

[9] Bryce Covert, *A Bail Reform Tool Intended To Curb Mass Incarceration Has Only Replicated Biases in the Criminal Justice System*, THE INTERCEPT (Jul. 12, 2020), https://theintercept.com/2020/07/12/risk-assessment-tools-bail-reform/; Laurel Eckhouse et al., *Layers of Bias*, SAGE PUBLICATIONS (Nov. 23, 2018), https://journals.sagepub.com/doi/10.1177/0093854818811379, p.196 (collecting studies reflecting that "people of color, especially Black people, are more likely to be arrested than Whites for the exact same behavior").

[10] Sandra G. Mayson, *Bias In, Bias Out*, YALE LAW JOURNAL (Sept. 28, 2018), https://www.yalelawjournal.org/pdf/Mayson_p5g2tz2m.pdf, n. 23; Laurel Eckhouse et al., *Layers of Bias*, SAGE PUBLICATIONS (Nov. 23, 2018), https://journals.sagepub.com/doi/10.1177/0093854818811379

[11] Ben Green, *The False Promise of Risk Assessments: Epistemic Reform and the Limits of Fairness*, HARVARD UNIVERSITY (Jan. 2020), *https://scholar.harvard.edu/files/bgreen/files/20-fat-risk.pdf;* Bryce Covert, *A Bail Reform Tool Intended To Curb Mass Incarceration Has Only Replicated Biases in the Criminal Justice System*, THE INTERCEPT (Jul. 12, 2020), https://theintercept.com/2020/07/12/risk-assessment-tools-bail-reform/

[12] *Updated Position on Risk Assessment Tools*, PRETRIAL JUSTICE INSTITUTE (Feb. 7, 2020), https://www.pretrial.org/wp-content/uploads/Risk-Statement-PJI-2020.pdf

[13] Victoria A. Terranova and Kyle C. Ward, *Colorado Pretrial Assessment Tool Validation Study Final Report*, UNIVERSITY OF NORTHERN COLORADO (Jul. 1, 2020), https://www.nacdl.org/getattachment/18510570-e0eb-4d40-b737-5aafb30c1085/terranovaward_cpat-validation-study_final-report.pdf, Table 22, p. 50 and Table 25, p. 52, which both reflect a similar base rate across Black and White defendants; Table 31, p. 58 also reflects equal base rates at 35-36%, but the underlying dataset appears problematic.

[14] In situations where base rates are not equal between groups, and we believed these base rate differences were justified (eg. if we didn't think inherent bias exists in the outcome variable itself such as re-arrest rates driven by over policing), Black people being placed in higher risk categories more often than White people does not *per se* guarantee disparities in false positive rates. In a situation where base rates are equal between groups, more people in higher risk groups implies disparities in false positives. See Sorelle A. Friedler et al., *On the (im)possibility of fairness*, ARXIV (Sept. 23, 2016), https://arxiv.org/pdf/1609.07236.pdf for an instructive explainer on challenges in achieving fairness. See Richard Berk et al., *Fairness in Criminal Justice Risk Assessments: The State of the Art*, ARXIV (May 30, 2017), https://arxiv.org/pdf/1703.09207.pdf for a criminal justice specific explainer that sites the impossibility of fairness when base rates are not equal -— a situation which according to the UNC study we are not facing here

[15] Victoria A. Terranova and Kyle C. Ward, *Colorado Pretrial Assessment Tool Validation Study Final Report*, UNIVERSITY OF NORTHERN COLORADO (Jul. 1, 2020), https://www.nacdl.org/getattachment/18510570-e0eb-4d40-b737-5aafb30c1085/terranovaward_cpat-validation-study_final-report.pdf, Appendix J, p. 95 which reflects the only CPAT evaluation of disparate impact in the report which we could identify

[16] *Id.*, Appendix J, p. 95 which reflects the only CPAT evaluation of disparate impact in the report which we could identify

[17] Department of Housing and Urban Development, *HUD 2019 Continuum of Care Homeless Assistance Programs Homeless Populations and Subpopulations*, DEPARTMENT OF HOUSING AND URBAN DEVELOPMENT (Sep. 20, 2019), https://files.hudexchange.info/reports/published/CoC_PopSub_NatlTerrDC_2019.pdf, which suggest ~1 in 5 homeless have a "severe mental illness", nearly 1 in 5 with "substance abuse problems" and 1 in 50 with HIV/AIDS

[18] *See, e.g.*, *Bearden v. Georgia*, 461 U.S. 660, 672-73 (1983), (finding that revoking probation and thereby "depriv[ing] the probationer of his conditional freedom simply because, through no fault of his own he cannot pay [a] fine . . . would be contrary to the fundamental fairness required by the Fourteenth Amendment"); *Tate v. Short*, 401 U.S. 395, 398 (1971) ("[T]he Constitution prohibits the State from imposing a fine as a sentence and then automatically converting it into a jail term solely because the defendant is indigent and cannot forthwith pay the fine in

full."); *William v. Illinois*, 399 U.S. 235, 241-42 (1970), (striking down the "invidious discrimination" of incarcerating a person beyond the statutory maximum term when he could not pay the imposed fine).

[19] Victoria A. Terranova and Kyle C. Ward, *Colorado Pretrial Assessment Tool Validation Study Final Report*, UNIVERSITY OF NORTHERN COLORADO (Jul. 1, 2020), https://www.nacdl.org/getattachment/18510570-e0eb-4d40-b737-5aafb30c1085/terranovaward_cpat-validation-study_final-report.pdf, Table 22 p. 50 and Appendix I, p. 93. Table 31, p. 58 suggests no disparities on a sub-sample of the test dataset, but there are reasons to be concerned about the test dataset's validity. The authors of the UNC study use multiple different datasets in their study including a retroactive dataset from 10 counties spanning 2015-2016, and a pilot set from the testing of CPAT-R in 7 of the 10 counties that lasted for 3 months in 2018-2019. They also often split these further into "validation", "testing" and "full sample" without clear documentation about which is from the validation set (label as "retroactive" here) and the test set (label as "pilot" set here). Unfortunately, the retroactive set and pilot set appear wildly different in many concerning ways — in the retroactive set 20.23% of people pretrial had a high consequence failure to appear and only 7.29% did in the pilot set. More concerning, in the pilot set *most* people's failures to appear were "no consequence" failures. Numerous other discrepancies appear to exist between the retroactive and pilot data and can be detected by comparing Table 16 and Table 4. All analyses are therefore colored by sample biases that may exist in each of these datasets (and both of these datasets already contain numerous sampling problems we cannot control including a lack of information about what would have happened to those who were jailed pretrial had they been released). We do not have sufficient information to appreciate which is "most correct."

[20] *Id.* p.42 ("Stakeholders across counties anecdotally advised that FTA's recorded in court records were often 'un-willful.' Un-willful FTA's were described as those that occur at no fault of the pretrial defendant.").

[21] Since the variable appears twice, a defendant immediately gets six points for failing to appear at a single court date, regardless of the cause.

[22] *Id.* Appendix I p. 93

[23] *See, e.g., U.S. v. Salerno*, 481 U.S. 739, 741 (1987), (upholding 1964 Bail Reform Act, which "narrowly" permits pretrial detention, after appropriate procedural safeguards, for "individuals who have been arrested for a specific category of extremely serious offenses"); *Reynolds v. U.S.*, 80 S.Ct. 30, 32 (1959), ("The purpose of bail is to insure the defendant's appearance and submission to the judgment of the court.")

[24] This is a problem inherent to risk assessments that CPAT cannot escape. For an overview, see eg Brandon Buskey and Andrea Woods, *Making Sense of Pretrial Risk Assessments*, NATIONAL ASSOCIATION OF CRIMINAL DEFENSE LAWYERS (Jun. 2018), https://www.nacdl.org/Article/June2018-MakingSenseofPretrialRiskAsses

[25] *Memorandum Addendum re Public Safety-Related Data*, COLORADO DIVISION OF CRIMINAL JUSTICE (Aug. 21, 2020)

[26] We know of one pretrial risk assessment which adds a flag for predicted "violent arrests" pretrial, known as the NVCA flag. However, only between 8.6% to 11% of those who receive this flag actually go on to be arrested for a violent offense. See, e.g. Laura and John Arnold Foundation, *Results from First Six Months of the Public Safety Assessment - Court in Kentucky, 3*, THE ARNOLD FOUNDATION (2014), http://www.arnoldfoundation.org/wp-content/uploads/2014/02/PSA-Court-Kentucky-6-Month-Report.pdf (indicating that the NVCA is associated with an 8.6 percent likelihood of arrest for a violent charge); Alexander Shalom et al., *The New Jersey Pretrial Justice Manual*, NATIONAL ASSOCIATION OF CRIMINAL DEFENSE LAWYERS (2016), https://www.nacdl.org/NJPretria (indicating that, in New Jersey, the NVCA is associated with an 11 percent likelihood of arrest for a violent charge).

[27] Victoria A. Terranova and Kyle C. Ward, *Colorado Pretrial Assessment Tool Validation Study Final Report*, UNIVERSITY OF NORTHERN COLORADO (Jul. 1, 2020), https://www.nacdl.org/getattachment/18510570-e0eb-4d40-b737-5aafb30c1085/terranovaward_cpat-validation-study_final-report.pdf, Table 4, p. 18. Of course, an arrest is not a conviction. We can be certain that of those charged with a violent offense, even less were convicted for violence.

[28] *Id.* Table 5 p. 19 which provides a confidence interval for AUC of violent arrests of 45-63 and accurately claims the following: "An AUC score of .50 or lower means that the risk assessment score predicts pretrial outcomes no better than chance".

[29] *Id.* p. 16.

[30] *Id.* p. 18, Table 4.

[31] Gini index is a mathematical transformation of AUC to convert the score to a percent scale such that the number represents a "% better than random guessing" where 0% represents random and 100% represents perfect estimates.

[32] Notably, both the CPAT and CPAT-R produce a single risk score for re-arrest and failure to appear — and the only reason CPAT-R "improves" upon CPAT is its over-emphasis on pretrial failures. The mixing of re-arrest and failure to appear into one outcome variable drives the entire improvement in AUC between the CPAT-R and CPAT — CPAT-R does a better job of predicting single FTA than CPAT but is otherwise equivalent. Given how very different re-arrests and failures to appear are, it's unreasonable to collapse these into a single risk score and again makes people seem much riskier than they are. Numerous studies suggest, especially for failures to appear, that interventions outside of caging people can dramatically reduce FTAs. See recent results such as Alissa Fishbane et al., *Behavioral nudges reduce failure to appear for court*, AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE (Oct. 8, 2020), https://science.sciencemag.org/content/early/2020/10/07/science.abb6591which reduced FTAs by 13-21% simply by improving court appearance forms.

[33] AUC is a metric that attempts to understand accuracy across every possible score of a risk assessment instrument — it essentially sorts our dataset and evaluates our ability to predict failures above/below each score. However, in reality, decisions are not made across a continuum, they are made at specific cut points which tool creators decide after the fact and without input from decision makers or those impacted by those decisions. In this ranking exercise, a model can look "good" because it ranks the very lowest risk category individuals — telling us with even greater certainty that people we already think will succeed, indeed will succeed — well even as it provides no insight into risks for those at the highest end of the ranking.

[34] Victoria A. Terranova and Kyle C. Ward, *Colorado Pretrial Assessment Tool Validation Study Final Report*, UNIVERSITY OF NORTHERN COLORADO (Jul. 1, 2020), https://www.nacdl.org/getattachment/18510570-e0eb-4d40-b737-5aafb30c1085/terranovaward_cpat-validation-study_final-report.pdf, Page 8 ("Pretrial assessment tools with fair performance have AUC scores ranging from .55-.63, good .64-.70 and excellent is .71 or higher.").

[35] *Id.* Table 9 Page 22 provides a direct comparison between CPAT and CPAT-R with slightly better AUC for the latter

[36] *Id.* For the threshold based accuracy metrics, which are the metrics that matter to us since they impact decision making in the courts, the UNC study provides *very* little information beyond the limited tables available in the paper (unlabeled table Page 12, Table 2 Page 14, unlabeled table Page 60). Critical information we are lacking for proper evaluation includes *how* the thresholds were chosen, the population of people in each of the datasets in each of the four categories, confidence intervals around these accuracy metrics, statistics about the secondary metrics at the 4 decision thresholds, and fairness/disparate impact metrics across each of the 4 thresholds for both CPAT and CPAT-R.

[37] The study leaves gaping holes in our understanding of both CPAT and CPAT-R in practice, most notably we have no understanding of judicial overrides and whether or not the CPAT will exacerbate existing biases in the courts. In other studies done throughout the country, one of the

biggest challenges in understanding whether risk assessments will truly do any good — and no harm — is evaluating how judges will interact with risk assessments (see earlier citations). In the qualitative section, the authors note this potential problem, but are more focused on "buy-in" than understanding what the impact of the decision will be. In many states, judicial overrides have dramatically *increased* existing biases. In others, they have at least nullified any potential decarceration effects from the use of risk assessments. Other studies aside, we have no idea what a "category 4" means to a judge, besides highest risk. Do we think judges know this means less than half of these people will fail to appear, only once, and even less will be re-arrested for anything, no matter how minor?