



NMCD Classification Instrument Validation

Summary of Findings, June 2024 Revision

Center for Applied Research and Analysis | University of New Mexico | April 2026

6,594 Unique individuals studied	22,193 Classification events	18 + 18 months Symmetric study window
--	--	---

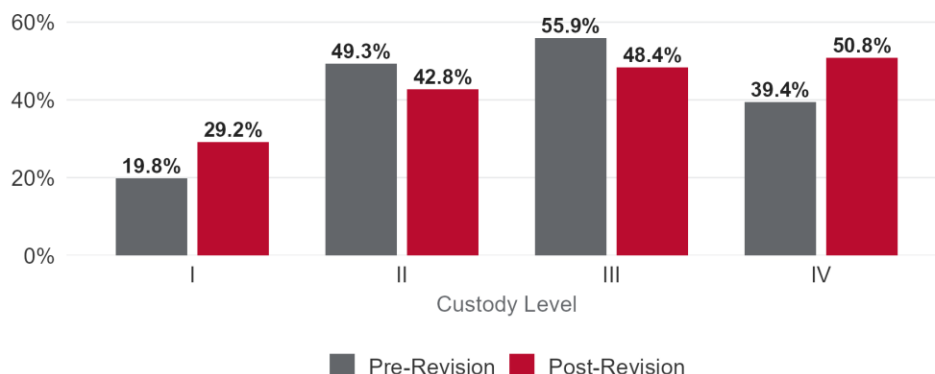
On June 3, 2024, NMCD revised its external custody classification scoring tools, modifying item weights and custody-level thresholds on both the initial and reclassification tools. This brief summarizes our validation study evaluating whether the revision is associated with changes in misconduct outcomes and classification accuracy, using 18 months of pre-revision and 18 months of post-revision data. *Full technical report available from CARA.*

What the revision did at the instrument level

- **Overall misconduct did not rise after the revision.** After adjusting for age, sex, score, and facility, the odds of any misconduct were about 25% lower post-revision than pre-revision (OR = 0.748, 95% CI [0.669, 0.837]), with the same direction for serious and serious violent outcomes. A within-person comparison of 750 individuals classified in both periods points the same way (56.5% misconduct rate before, 48.0% after, $p = .001$). These are associations, not proof of cause. Because several policy changes (notably a relaxation of mandatory override rules) took effect on the same day and could plausibly have pushed misconduct upward, the absence of a broad-based increase is itself reassuring.
- **The reclassification tool got better at separating higher- from lower-risk individuals.** Its ability to predict misconduct, measured by AUC (where 0.50 is chance and higher values mean better separation), rose from 0.561 to 0.624 for any misconduct, from 0.559 to 0.604 for serious misconduct, and from 0.583 to 0.606 for serious violent misconduct. On the Desmarais and Singh (2013) scale, the tool moved from the lower end of the “fair” range toward the “good” threshold of 0.64. None of the within-tool differences reach conventional statistical significance on their own, but the direction is consistent across all three outcomes.
- **The custody-level gradient now matches the tool’s design.** Before the revision, Level IV misconduct (39.4%) was actually lower than Level III misconduct (55.9%), which is the opposite of what the instrument is intended to produce. After the revision, Level IV (50.8%) now sits above Level III (48.4%), so observed misconduct rises stepwise with custody level, as intended.
- **Mandatory overrides fell sharply.** System-wide, mandatory overrides dropped from 50.2% to 38.3%. On the initial tool specifically, they fell from 65.0% to 42.7%. A larger share of placements now follows the scored recommendation rather than a rules-based adjustment.

The Level III/IV inversion has been resolved

183-day any-misconduct rate by custody level, non-overridden cohort



What drives prediction on each tool

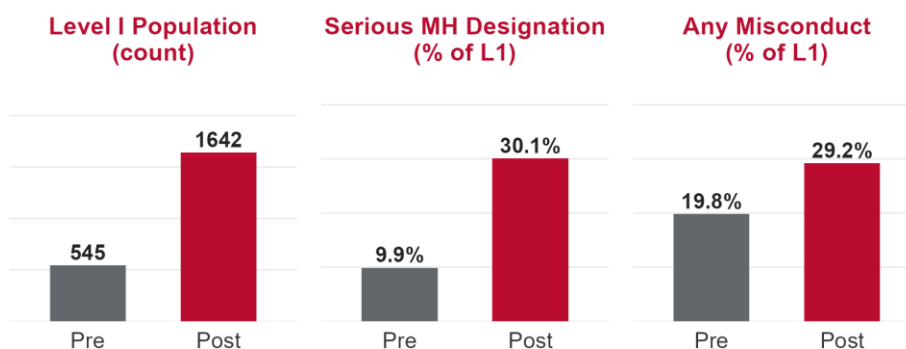
The two tools rely on different mixes of predictors. On the initial classification tool, age is doing nearly all of the predictive work; other scored items (offense severity, prior felonies, escape history) add little predictive power once age is accounted for. On the reclassification tool, the signal is spread more evenly across several items, led by institutional disciplinary history, age, and institutional violence history, with meaningful contributions from program participation, prior felonies, and conviction severity. In practical terms, the initial tool functions largely as an age-based instrument at intake, while the reclassification tool is able to draw on observable in-custody behavior once it becomes available.

Where the post-revision pattern raises questions

- **Level I misconduct rose from 19.8% to 29.2%.** This reflects a change in who is at Level I, not a change in behavior among individuals already there. The post-revision Level I population is about three times larger and includes a much higher share of people with serious mental-health designations (9.9% → 30.1%) and younger individuals, both of whom carry elevated misconduct risk. Whether Level I housing, programming, and supervision are well-matched to this new population is a question the current data cannot answer and is best addressed in the next validation cycle using facility-level operational data.

Level I composition shifted post-revision

The new Level I cohort is larger, carries more mental-health burden, and shows higher misconduct



- **Discretionary overrides roughly doubled, from 10.7% to 22.4%.** Rules-based overrides went down, but overrides initiated by staff judgment rose by a similar amount. Periodic audits of override

decisions are warranted, particularly downward overrides into Level I, which are linked to substantially higher misconduct than non-overridden Level I placements.

- **Misconduct is highly concentrated in a small subset of individuals.** About 10% of non-overridden individuals account for roughly 55% of all guilty misconduct findings, and about 20% account for 77–80%. Targeted interventions aimed at these high-frequency event-generators would, in principle, reduce a majority of the total misconduct load.

Recommendations

- Maintain the revised scoring instrument. The instrument-level improvements are material, and none of the adjusted analyses indicate a post-revision increase in misconduct.
- Align Level I programming and service capacity with the post-revision population. The Level I compositional shift toward younger individuals and those with serious mental-health designations warrants expanded mental-health programming and crisis intervention capacity at Level I facilities, along with targeted orientation and substance-use programming for young initial-classification Level I individuals.
- Pair future validation cycles with facility-level operational data. Level I questions about service capacity, housing density, and programming availability are directly addressable with operational data.
- Review discretionary downward-override practices into Level I. Downward-overridden Level I inmates have roughly double the any-misconduct rate of non-overridden Level I inmates.
- Continue validation at regular intervals. As the classified population and facility operations evolve, periodic revalidation ensures the instrument remains calibrated.

Methods in brief

Cohort design using 18 months of pre-revision data (December 2022 to June 2024) and 18 months of post-revision data (June 2024 to December 2025). Each classified individual is followed for 183 days (about six months) for any misconduct event, with follow-up ending early if they are reclassified. Most analyses focus on the non-overridden cohort (approximately 9,600 classifications from 3,573 people), where the scored level and the final custody level agree. Methods include adjusted logistic regression, Cox survival analysis, interrupted time series with placebo comparisons, within-person paired comparisons, propensity-weighted sensitivity checks, regression-discontinuity analyses at level boundaries, and concentration (Pareto and Lorenz) analyses. All inferential models use cluster-robust standard errors clustered on the individual.

Alexander Severson, Ph.D. | Director and Senior Research Scientist II | CARA | Institute for Social Research | University of New Mexico