The Development and Validation of the Minnesota Severe and Frequent Estimate for Discipline (MnSafeD)

Grant Duwe
Director, Research and Evaluation
Email: grant.duwe@state.mn.us

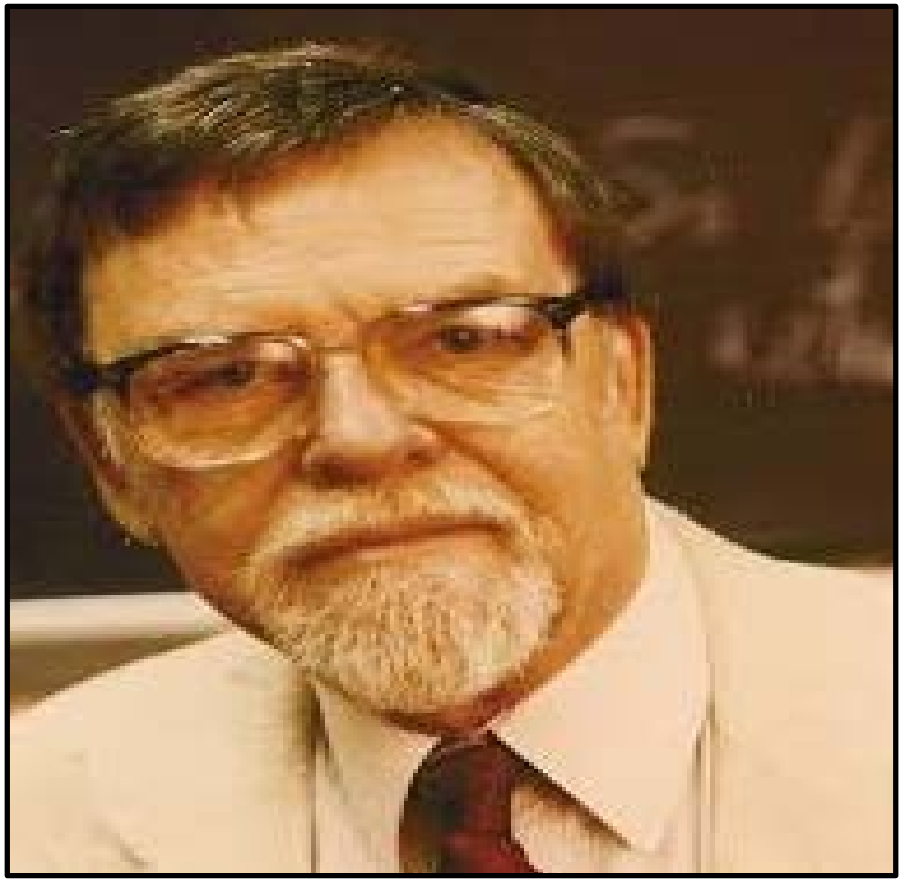
September 2019

Overview

- Background
  - Professional vs. Actuarial Judgment
  - Customization
  - Automation
  - Bias
- Minnesota Severe and Frequent Estimate for Discipline (MnSafeD)
  - Development
  - Validation
- Predictive Performance Results
- Next Steps

Paul Meehl: Clinical Psychologist at U of M

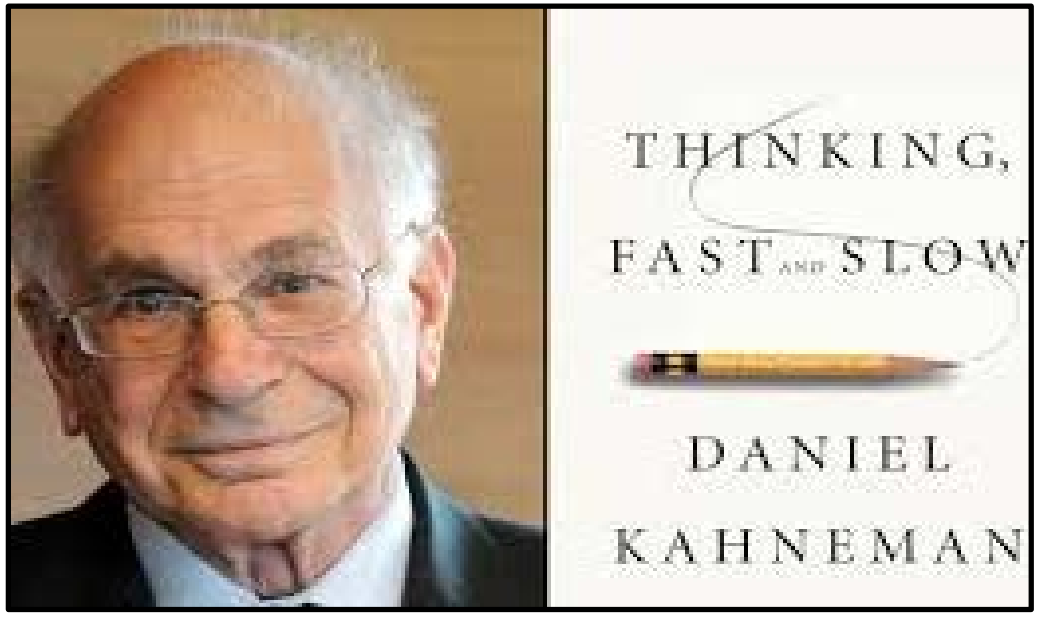
- Book in 1954: Clinical vs. Statistical Prediction
- Found that "mechanical" (formal, algorithmic) prediction outperformed "clinical" judgment (informal and subjective)
- Mechanical prediction is more reliable→ consistent
- In the more than 70 years since Meehl's book, research from a variety of fields has consistently confirmed that statistical prediction outperforms clinical judgment



➤ Professor of Psychology at U of M from 1945 to 2003

Daniel Kahneman and Amos Tversky

- Kahneman and Tversky research on cognitive biases in decision-making
  - Why statistical prediction outperforms professional judgment
- Examples
  - Confirmation bias: discredit unsupportive info
  - Anchoring: excessive weight to unimportant characteristics
  - Familiarity/availability: situations seem similar
  - Base rate bias: favor specific info about a case vs. general info about group



- Currently Professor at Princeton University
- Won the 2002 Nobel Prize in Economics for work on decision-making

Research on Use of "Expert" Judgment

- Fields outside corrections that make risk assessment decisions
  - Health care, financial lending, insurance, stock trading
  - Evidence consistently shows that algorithms perform better than "expert" opinion or professional judgment
    - This is why all of these fields now rely mostly on algorithms/statistical prediction to make risk assessment decisions (process is often automated)
      - More valid, reliable, objective, efficient and cost-effective
- Corrections → predicting who will recidivate
  - General recidivism for correctional populations
    - Professional overrides led to reduced predictive performance
      - Wormith et al. (2012)
      - McCafferty (2017)

Prior Research on Customization

- Not much research has explicitly addressed this issue
  - Customized vs. Global, "off the shelf"
- But here's what we know:
  - A few studies suggest local instruments likely have better performance than assessments developed on other correctional populations
- Example: Level of Service (LS) family of tools (LSI-R and LS/CMI)
  - Most widely-used assessment for general recidivism
  - LS tools = developed and validated on Canadian correctional populations
  - Meta-analysis of LS validation studies (Olver et al., 2014)
    - Best performance for LS tools → studies on Canadian offender populations
    - Worst performance for LS tools → studies on U.S. offender populations
      - Validation research on the LSI-R for MN prisoners confirms this
      - LSI-R = relatively poor performance in predicting recidivism for MN prisoners

The "Home-Field Advantage"

- Study by Duwe and Rocque (2018): MnSOST-3 outperformed Static-99 on MN sex offender population
  - Static-99: developed on SO population from Canada/UK
  - MnSOST-3: developed on MN sex offender population
- There is a home-field advantage to risk assessment
  - Home-grown assessments will (all else being equal) likely outperform assessments developed elsewhere
  - Common Practice—what usually happens
    - Use assessment developed/validated on another correctional population
      - Assume assessment will perform just as well on own population
      - This is not a safe assumption to make
  - What should happen → An assessment's performance should be evaluated/tested before it is used to help inform decisions

Impact of Using Automated Scoring Method

- Duwe and Rocque (2017) study in *Criminology & Public Policy*
  - Examined effects of automated risk assessment on reliability, predictive validity and return on investment (ROI)
- Minnesota DOC began using MnSTARR in 2013
  - Gender-specific, manually-scored assessment risk for multiple types of recidivism
    - Felony, non-violent, violent and sexual offending
    - Static and dynamic items
  - Average = 35 minutes to score (by prison caseworkers)
- MnDOC → Began using MnSTARR 2.0 in 2016
  - Similar to original MnSTARR but…
    - Fully-automated assessment (prison staff do not score it)
      - Overnight batch process and/or generated by caseworker (10-15 seconds to run)
    - About 2X the number of items (nearly 50 total)

Results from Duwe and Rocque (2017) Study

- Automation eliminates inter-rater disagreement
  - Every assessment is scored the same way (removes layer of error)
  - Doesn't mean data are flawless
- Increased reliability → Better predictive performance
  - As reliability got worse in manual assessments, so did the predictive performance
    - Cases w/ more inter-rater disagreement = worse predictive performance
- Investment/Cost = $135,000 to automate (a one-time cost)
- Return/Benefits = MnDOC staff time saved from automation
  - Monetized staff time = salary/benefits for prison caseworkers
  - Automation = major increase in assessment capacity
- Benefit/Cost Estimate after:
  - Year 1 = $452,108; ROI = $4.35 **(Actual = $955,990; ROI = $8.08)**
  - Year 2 = $1.04 million; ROI = $8.70 **(Actual = $1.8 million; ROI = $13.32)**
  - Year 5 = $2.8 million; ROI = $21.74

Bias in Risk Assessment

- ProPublica → Use of COMPAS in Florida
  - Allegations of racial bias
- Canada → performance for indigenous population
- A lot of confusion/misunderstanding
  - Risk Assessments used in a lot of different ways
  - Alternative?
    - Human/Professional Judgment = more biased
- Imperative to test for bias
  - Evaluate performance among sub-populations
  - Beyond this, not much guidance (yet)
    - A difference in performance does not equate to bias
      - Example: AUC of 0.90 versus AUC of 0.85

MnDOC Current Classification System

- Late 1990s → MnDOC implemented a classification assessment
  - Received technical assistance from NIC (like a lot of other states)
- MnDOC Classification Assessment
  - Scored manually by staff
    - Conduct a file/database review
  - 6 Items
    - Current offense
    - History of assault
    - Institutional adjustment
    - History of escape
    - Age
    - Custody level at most recent release
  - Uses a simple, summative weighting scheme (Burgess)
- Parole violator admissions = not reassessed
- Never validated…until now

What's the MnSafeD?

- A fully-automated, gender-specific classification assessment that predicts severe and frequent misconduct for individuals in prison on a recurring, semi-annual basis
  - Classification assessments used to help make security/custody level decisions for those in prison
- Developed on sample of 39,355 releases from Minnesota prisons (2006-2011)
  - 35,506 males
  - 3,849 females
- Used bootstrap resampling, k-fold and split-population methods to select predictors and validate/test predictive performance
- Used multiple metrics to evaluate predictive performance

Predicting Prison Misconduct

- MnSafeD predicts "severe and frequent misconduct"
  - Multiple discipline convictions and/or violent/assaultive misconduct within a six-month period
    - About 10% of Minnesota's prison population
- Why not just predict all misconduct?
  - Nearly one-third of MN inmates have at least one discipline conviction (DC)
    - Attempting to predict who will have at least one DC = not helpful in managing risk
- Insight from career criminal literature
  - Small # of prolific offenders responsible for a lot of crime
  - Same is true for misconduct
    - 10% of MN prisoners = 70% of all DCs, 80% of seg DCs and 100% of violent DCs (males)
      - Compromise safety for staff and other inmates
    - Predictors of recidivism and prison misconduct = a lot of overlap

Other Design Assumptions

- Gender-specific
  - Potential gender differences in risk and protective factors
  - Males and females also housed in different facilities
    - Misconduct can be influenced by facility-level factors
- Fully-Automated Scoring Method
  - More reliable, valid, efficient and cost-effective than a manual scoring method
  - MnSafeD leverages work on MnSTARR 2.0
    - Fully-automated recidivism risk assessment used by MnDOC since November 2016
- Assessment predicts SFM at intake and reassesses every 6 months thereafter
  - This is how MnDOC uses its current classification assessment
    - Current classification assessment = predictive performance never evaluated
    - Based on NIC model from late 1990s (like a lot of state DOC's)

Model Development and Validation

- Regularized logistic regression = classification algorithm
  - "Shrinks" large coefficients to reduce overfitting
- Used bootstrap resampling method to help identify significant, robust predictors
  - $P < .05$ in at least 70% of 1K bootstrap samples
- Validation
  - Split samples into training (2006-2009 releases) and test (2010-2011 releases) sets; also used additional test set (2017 admissions)
  - Using 10-fold CV, varied ridge estimator value on training set data to help identify the best performing model
  - Best models were then applied to test sets to evaluate predictive performance
- Performance Metrics
  - ACC, AUC, H, PRC, RMSE, SAR and SHARP
    - Focus on AUC (for this presentation)

Dataset

- Predicted Outcome = SFM within a six-month window or release
  - Multiple discipline convictions and/or violent/assaultive misconduct within a six-month period
- Predictors (similar to those used for MnSTARR 2.0)
  - Criminal history
    - Type/severity of offenses, specialization in specific offenses (violent, felony, drug, etc.)
  - Offense type (index)
  - Prison admission type
  - Suicidal tendencies
  - Security threat group (gang affiliation)
  - Demographics→ age at release, marital status
- Main difference in predictors (between MnSTARR & MnSafeD)
  - Also considered prior prison misconduct (for those in prison previously)
  - Incorporated recent prison data for reassessments
    - Prison misconduct (frequency and severity)
    - Involvement in prison programming
      - UI status = unauthorized idle
- Data split up in 6-month intervals (per inmate)

Example: Male Prisoner Dataset

	Training Set (N)	Test Set (N)
Intake	23,838	11,668
Intake (2017 test set)	23,838	3,468
6-Month Reassessment	12,481	6,875
6-Month Reassessment (2017)	12,481	735
12-Month Reassessment	7,778	4,468
18-Month Reassessment	5,247	2,833
24-Month Reassessment	3,745	1,994
30-Month Reassessment	2,724	1,447
36-Month Reassessment	1,886	1,032
42-Month Reassessment	1,365	767

Predictive Performance Results for Female Test Set

	Current Classification (AUC)	MnSafeD (AUC)	Training Set N	Test Set N
Intake	0.628	0.759	2,546	1,303
Intake (2017 test set)	0.607	0.731	2,546	710
6-Month Reassessment	0.655	0.854	1,076	592
6-Month Reassessment (2017)	0.650	0.922	1,076	177
12-Month Reassessment	0.694	0.909	562	352
18-Month Reassessment	0.681	0.819	312	211
Overall Average	0.653	0.832		

- AUC "rule of thumb"
  - >= 0.90 ' "A"
  - 0.80-0.89 = "B"
  - 0.70-0.79 = "C"
  - 0.60-0.69 = "D"
  - < 0.60 = "F"

Predictive Performance Results for Male Test Set

	Current Classification (AUC)	MnSafeD (AUC)	Training Set N	Test Set N
Intake	0.632	0.768	23,838	11,668
Intake (2017 test set)	0.617	0.747	23,838	3,468
6-Month Reassessment	0.665	0.828	12,481	6,875
6-Month Reassessment (2017)	0.650	0.800	12,481	735
12-Month Reassessment	0.674	0.857	7,778	4,468
18-Month Reassessment	0.674	0.876	5,247	2,833
24-Month Reassessment	0.690	0.884	3,745	1,994
30-Month Reassessment	0.666	0.871	2,724	1,447
36-Month Reassessment	0.688	0.888	1,886	1,032
42-Month Reassessment	0.697	0.840	1,365	767
Overall Average	0.665	0.836		

Explaining the Results

- MnSafeD = high level of predictive performance
  - Better than what's usually observed for recidivism, including MnSTARR (recidivism risk assessment for MN prisoners)
- Why?
  - Predictive performance advantages:
    - Customized to MN population = "home field advantage"
    - Uses automated scoring = more reliable (no inter-rater disagreement)
    - Classification algorithm: RLR > Burgess methods
- Better than MnSTARR 2.0
  - Recent behavioral indicators = influential in predicting SFM
    - Severity and frequency of prison misconduct in last 6 months or since most recent admission to prison
    - UI status (no programming) in last 6 months

Next Steps

- MnSafeD = MnDOC new classification assessment
  - MnDOC IT currently working on implementing the MnSafeD
- MnSafeD will be used to help determine custody-level placement
- Custody-level assignment is important
  - But should it be the only way a classification assessment is used?

Making the Case for Front-Loading

- Programming often "back-loaded" closer to time of release
  - There's good reason for this → better recidivism outcomes
- Improving institutional safety = more than just custody-level placement
- Front-loading programming
  - …at least for those at high-risk of SFM
    - Deliver programming to those at high risk of SFM shortly after intake/beginning of confinement
      - Example: immediately prioritize those at highest risk of SFM (top 5 percent) for an intervention (e.g., cognitive-behavioral therapy) at the beginning of confinement
    - Front-loading may not only reduce misconduct but also increase dosage
      - Greater dosage = better recidivism outcomes

Final Thoughts

- A lot of prison systems still use what are, by now, outdated classification assessments
  - How are these performing?
- MnSafeD represents one approach
  - Not designed to be a one-size-fits-all solution
  - Some of it may be worth replicating in the event prison systems (or jail systems) upgrade their classification assessments
- MnSafeD study will be published in *The Prison Journal*
  - Citation: Duwe, G. (forthcoming). The development and validation of a prison classification system designed to predict severe and frequent misconduct. *The Prison Journal*.