Recently completed study with Michael Rocque (Bates College) to be published in *Criminology & Public Policy* that:

- Compares manual scoring method with an automated scoring process
  - Manual = correctional staff enter values for each item on a risk assessment tool by hand
  - Automated = items on the instrument are auto-populated (does not require staff to enter values for these items)
- Compares the reliability (i.e., consistency with which an instrument can be scored across different raters) of the manual and automated processes
  - Examined the impact that inter-rater disagreement had on performance in predicting recidivism
- Estimates the return on investment (ROI) for fully automating the Minnesota Screening Tool Assessing Recidivism Risk (MnSTARR)
Background

- Minnesota DOC began using the MnSTARR in April 2013
  - Study on development of MnSTARR published in *Criminal Justice Policy Review*
  - MnSTARR is strictly a recidivism risk assessment tool
    - Does not assess needs (i.e., identify which interventions would be most appropriate for an offender)
    - But it does include dynamic items that measure whether needs areas have improved or grown worse while in prison
      - Worse = prison misconduct, involvement in gang/security threat group
      - Better = earned educational degree/certificate, completed treatment, get visited in prison
  - Study found MnSTARR significantly outperformed LSI-R in predicting recidivism, so…
    - MnSTARR used to assess recidivism risk
    - LSI-R (and now the LS/CMI) used to assess needs for higher-risk offenders
More on the MnSTARR

• Separate scales/items for male and female offenders
• Assesses risk (over a 4-year follow-up period) for the following types of recidivism
  • Felony
  • Non-violent
  • Non-sexual violent
  • Sexual offending (males only)
    • First-time and repeat sexual offending
• Offenders placed in 1 of 4 risk levels:
  • Very high, High, Medium and Low
• Administered (manually for the most part) twice
  • Once at intake and one more time prior to release
• Used to prioritize offenders for programming (institution and community) and determining level of post-release supervision
The MnSTARR 2.0

- In late 2016, Minnesota will implement the MnSTARR 2.0, which is:
  - A fully-automated recidivism risk assessment tool
    - Implications from Automation
      - More items can be included since they don’t have to be scored by hand
      - Substantially increases assessment capacity
        - Currently, only offenders w/ 6 months or more get MnSTARR
        - Under MnSTARR 2.0, everyone gets assessed for risk
      - Performance—reliability, predictive validity and ROI
      - Staff time saved
        - Average time to score the MnSTARR = 35 minutes
        - Average time to score the MnSTARR 2.0 = 7 seconds
Reliability of Risk/Needs Assessment

- Inter-rater reliability (IRR): consistency between raters in scoring an instrument
  - Very important for manually-scored tools
- In theory, reliability is intertwined with validity
  - An unreliable tool (more inter-rater disagreement) will have lower predictive validity
- Few studies have examined IRR
  - None have examined relationship between IRR and predictive performance
- IRR study on LSI-R (Rocque & Plummer-Beale, 2014)
  - ICC (intraclass correlation coefficient) = 0.65
    - 0.60-0.75 = “good” and above 0.75 is “excellent”
Predictive Performance

• Nearly all of the research has focused on predictive performance
  How well does an instrument predict recidivism?

• Predictive validity has 3 main dimensions
  Predictive discrimination
  • How well does tool separate the recidivists from the non-recidivists?
  • Commonly-used predictive discrimination measures
    • Area under the Curve (AUC): 0.70 and above is adequate
    • Hand’s H-measure

• Predictive accuracy
  • How well does tool accurately classify offenders?
    • Recidivist with a recidivism probability > 50% = true positive
    • Non-recidivist with a recidivism probability < 50% = false positive
    • Accuracy (ACC) is a commonly-used measure

• Calibration
  • How well do predicted probabilities align with observed recidivism outcomes?
    • Root mean squared error (RMSE) and Brier score are some commonly-used measures

• Other measures of predictive performance include:
  • Precision, Recall, F-score, etc.
Prior Research on Predictive Performance

• Most has focused only on predictive discrimination
  • AUC being the most common measured used
  • Performance varies both within (depending on who’s doing the validating) and across tools

• Research has recently focused on impact of classification method on performance
  • Burgess = simple, summative method—most common among widely-used tools
  • Supervised learning algorithms: ranges from types of logistic regression models to “machine learning”
    • Supervised learning algorithms = superior performance over Burgess methodology
Current Study on Reliability and Predictive Validity

- Sample = 3,985 offenders (400 females and 3,585 males) released in 2014 who had been scored on MnSTARR
  - The assessments, scored by prison caseworkers, represent data from a manual process
  - Also scored these same offenders on the MnSTARR using an automated process
- Collected one-year recidivism data to evaluate predictive performance
- Used data from an older sample of offenders released from 2003-2010 to develop and validate models based on a one-year follow-up period
  - Applied these models to the manual and automated assessment data
IRR Results

- Compared the assessments scored manually with those scored by the automated process
  - Results show assessments by prison caseworkers were generally consistent with automated scoring process
    - All ICC values were in “excellent” range (0.75 and above)
  - Male Offenders
    - Average ICC = 0.86
      - High of 0.89 for general recidivism
      - Low of 0.81 for violent recidivism
  - Female Offender
    - Average ICC = 0.90
      - High of 0.94 for felony recidivism
      - Low of 0.81 for violent recidivism
IRR and Predictive Performance

- Despite “excellent” ICC values, manual assessments had inferior performance compared to automated process.

- Used 5 predictive performance measures:
  - AUC, H-measure, ACC, RMSE, and SAR.

- Compared to automated process, manual assessments had worse performance across all 5 performance measures:
  - For both male and female offenders.
  - For each recidivism measure evaluated.

- Example:
  - Average AUC for males:
    - 0.736 for automated and 0.714 for manual.
  - Average AUC for females:
    - 0.759 for automated and 0.730 for manual.
A Closer Look at IRR and Predictive Performance

- All ICC values in “excellent” range, so we arranged male offender data into quintiles:
  - Top 20%: ICC = 0.999
  - 61-80%: ICC = 0.990
  - 41-60%: ICC = 0.966
  - 21-40%: ICC = 0.916
  - Bottom 20%: ICC = 0.615 (similar to Rocque & Plummer-Beale, 2014)
- Little difference in predictive performance when ICC > 0.95
- Performance for manual assessments much worse for bottom 20% (greater disagreement with automated process)
  - Average difference in AUC = .046
  - Average difference in H-measure = .042
  - Similar to difference observed for classification method:
    - AUC/H-measure about .05-.07 higher for supervised learning algorithms compared to Burgess methodology
Could you put that in English/Practical Terms?

- Nearly 8,000 offenders released each year from Minnesota prisons
- Compared to an automated process, using a manual process with an ICC in the “good” range (reduces AUC/H by .04-.05) would result in:
  - More than 1,000 offenders being misclassified
    - False positives: offenders classified as “recidivists” who do not recidivate
    - False negatives: offenders classified as “non-recidivists” who recidivate
ROI for Automation

- **Investment/Cost** = $135,000 to automate the MnSTARR 2.0 (a one-time cost)
- **Return/Benefits** = Minnesota DOC staff time saved from automation
  - Monetized staff time = salary/benefits for prison caseworkers
  - Automation = 12,950 hours saved each year
    - Current assessments plus increased assessment capacity for short-stay offenders
- **Benefit/Cost Estimate after:**
  - Year 1 = $452,108; Ratio = $4.35
  - Year 5 = $2.8 million; Ratio = $21.74
  - Year 15 = $8.4 million; Ratio = $65.23
  - Community employment/job training = $43.26 (Aos and Drake, 2013)
Main Implications from Automation Study

- Inter-rater disagreement can have a major impact on predictive performance
  - Nearly as much as the classification method
- Thresholds for ICC = overly optimistic
  - If ICC between 0.60-0.74 is “good” but drops AUC by .05, it may not be so good after all
- Tentatively propose the following
  - 0.95 and above = excellent
  - 0.85-0.94 = good
  - 0.75-0.84 = adequate
  - Below 0.75 = poor
More Investment in Automation

• Manually-scored tools can still achieve adequate performance
  • But very costly in terms of time/money—ongoing training, ensuring quality, staff time
• Other industries that make risk assessment decisions (financial lending, insurance, healthcare) have shifted to automation
  • More objective
  • More reliable and valid
  • Greater efficiency and cost-effectiveness
Other Considerations

• Automated risk assessment saves time and money, which creates reinvestment opportunities
  • Minnesota DOC = greater focus on case planning
• Automation is not confined to risk assessment
  • Meredith (2014) developed automated needs assessment for Georgia DOC
  • Other options for increasing efficiency/performance of needs assessment process include:
    • Using computer-assisted survey software
    • Using more sophisticated supervised learning/machine learning methods
• Size of ROI will vary
  • To what extent does IT infrastructure need to be upgraded?
  • Economy of scale matters—bigger return for systems with larger populations