VERITAS UNITAS CARITAS
1842

VILLANOVA
UNIVERSITY

DEPARTMENT of SOCIOLOGY and CRIMINOLOGY

June 13, 2018

Good afternoon, your honor, members of the Pennsylvania Sentencing Commission, and fellow Pennsylvania citizens. I am a Professor of Sociology and Criminology at Villanova University who conducts quantitative research on a variety of issues related to crime and justice. The statistics and social science methods utilized in the Commission's many reports are quite familiar to me. I am here today because I feel that my experience in quantitative social science methods might be helpful for understanding the strengths and weaknesses of this proposed risk assessment instrument. My intention is to mostly limit my comments to issues of research design and statistical inference, as there are many other citizens here who are better qualified to discuss legal, political, ethical, and practical implementation issues.

Perhaps different from some others here today, I believe that predictive algorithms can be useful in certain criminal justice system contexts. In my opinion, such prognostic tools are most likely to be helpful when they provide very specific risk-related information that is known to counteract very specific human biases. Predicative algorithms are likely to be harmful when their goal is not linked to very specific human biases and the information presented to human decision makers is unbalanced.

Unfortunately, I believe the term unbalanced aptly describes how the current risk assessment proposal is fatally flawed. Today, I will present just a few examples of unevenness in the research design and application of statistical rules. I will ultimately argue that one cannot expect the public to have faith in this instrument's objectivity when the rules governing statistical inference and research design are applied inconsistently by its makers.
Before going further, I want to be clear on two points. First, as someone who has conducted quite a bit of quantitative social science research over the last 20 years, I have a sincere appreciation for how difficult this type of work is. Reviewing the Commission's many reports I was struck by just how much effort went into this endeavor and there were many pieces of the analysis that, as a technician, I could not help but marvel at. But, just like having many quality ingredients does not ensure a fine dinner (you need a good recipe), how the specific components are mixed together matters. Second, I do not mean to suggest that the unevenness in the Commission's analysis is a product of intentional biases. Often times, disparate outcomes are more about what decision makers fail to contemplate, rather than what they actually think.

With that said, consider the Commission's justifications for the decision to develop a risk scale for each Offense Gravity Score (OGS). Concerned that one size does not fit all, the Commission investigated whether there was empirical justification for developing nine separate risk assessment tools by the gravity of the initial offense. Unfortunately, that concern did not spark an analogous investigation of whether it made sense to differentiate recidivism by offense seriousness. Despite the Commission's own evidence that current and future offenses are often quite different in type and seriousness, as well as the Commission's mandate to specifically consider the threat to public safety, arrest-related recidivism is not differentiated by crime severity in the proposed instrument. Thus, the problem of lumping apples with oranges was deemed worthy of significant attention when considering current offenses, but for some reason, not future offenses.

Moreover, in justifying their approach, the Commission noted that "Development of a risk tool for all offenders as a group would potentially result in the less serious offenders having higher risk scores while more serious offenders would have lower risk scores." However, methodologically speaking, there is no reason to believe that the approach of creating OGS-specific scales will remedy this problem. So, for example, comparing men to men and women to women in terms of the amount of housework they do will not change the fact that women typically carry a heavier load than men. So, unsurprisingly, if you look at the final proposed instruments, the average risk scores for less serious offenders are still sometimes higher than those for more serious offenders. More troubling, offenders in the least serious offense category need only have a 54% predicted probability of committing any general offense to be labeled "High Risk," while offenders in the most serious offense category need to pass a 74% predicted probability threshold of committing the same offense to be similarly branded. Regardless of how you look at it, the instrument continues to assign higher risk designations to less serious offenders.

Another key justification for the Commission's decision to use OGS-specific scales was that such a move would lead to higher AUC values, with the value added achieving statistical significance for 3 of the 9 OGSs, and approaching significance for another. The Commission concluded from this evidence that the OGS-specific scales produced "modest improvement." As far as I can tell, this is the only part of the Commission's analysis where they demarcate a very low significance threshold of .10, or 1 out of 10.

In social science, there is considerable disagreement about the use, interpretation, and presentation of statistical significance tests. However, there is a consensus that is crucial for researchers to be consistent in the use, interpretation, and presentation of such tests throughout a particular project. One cannot cherry pick the cases when the rules of statistical inference apply and when they do not.

In the Commission's most recent racial impact analysis, where, rather than indicating degree of improvement, discerning statistically significant differences reflects negatively on the quality of the assessment tool, the Commission chose to take a different stance on statistical significance testing. In this case, the Commission states: "Following common practice, the results reported in Table 3 are based on a significance value of .001, which is appropriate with large sample sizes. The stricter standard is applied because in large sample sizes, trivial differences can attain statistical significance." However, one has to wonder why this stricter standard of 1 out of 1,000 was not applied earlier, especially in cases where the sample size was actually larger.[i] In order for something to be properly labeled a "standard," it must be consistently applied. Likewise, if a 2-3% estimated difference in AUC values constitutes a "modest improvement" that justifies the OGS-specific scales, in the context of racial disparity, it seems inappropriate to reference the same 2-3% AUC discrepancy as evidence of "similar accuracy."

In sum, I believe the current risk assessment proposal, despite the dedicated hard work that obviously went into it, is very uneven. In my testimony today I have focused on how the justifications for the decision to use the OGS-specific scales are problematic in a multitude of ways. First, if you care about the seriousness of the initial offense, it makes no sense to ignore the seriousness in how people might recidivate (and including information about crimes against a person in less than 1/10 of one percent of all cases obviously does not cover it). Second, OGS-specific scales do not solve one of the problems that motivated their creation: namely that those committing less serious offenses are more likely to be labeled high risk. Third, the parameters of what constitutes statistically and substantively significant differences for the justification of the OGS-specific scales are incongruent with the standards used elsewhere in the analyses. It is especially troubling that the Commission thought it appropriate to lower the bar for attaining statistical significance in the case of defending OGS-specific scales and raise the bar for determining what is statistically significant for uncovering racial disparities. Going forward, it is essential that the Commission place a greater emphasis on methodological consistency, as people will rightfully question the legitimacy of a system based on findings that have the appearance of being cherry picked. Finally, I implore the Commission to take a broader view of what constitutes a threat to public safety and to recognize that unwarranted social control is not simply “Cautious Error.” It is itself dangerous to individuals, families, and communities.

Thank you for your time.

Sincerely,

Lance Hannon

Lance Hannon, Ph.D.
Full Professor of Sociology and Criminology

[i] For example, with a sample size of 16,620 for OGS 5 in the black-white racial impact analysis, a probability value that was less than .05 but not less than .001 was deemed statistically insignificant, whereas with a sample size over 18,000 for OGS 3 in the OGS-specific vs. OGS-aggregated analysis, a probability value in that range achieved statistical significance.