THE IMPACT OF ERROR ON OFFENDER RISK CLASSIFICATION

By

Aaron K.T. Ho

A Dissertation submitted to the

Graduate School-Newark

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Criminal Justice

Written under the direction of

Dr. Todd Clear

And approved by

_____

_____

_____

_____

Newark, New Jersey

May, 2013

ABSTRACT

The Impact of Error on Offender Risk Classification

By Aaron K.T. Ho

Dissertation Director: Todd Clear

In criminal justice, offender risk classification seeks to divide individuals into different groups, normally so that varying levels of program treatment, custody, or supervision can be effectively and optimally allocated. The goal of effectively separating offenders based on prearranged criteria, however, is often thwarted by error problems, resulting in the misclassification of individuals. How the initial error problems eventually translate into final misclassification is not completely understood. Thus, the dissertation attempts to model the effects of error on the tolerance of offender risk classification instruments. Specifically, different properties and characteristics of classification devices are analyzed to understand their impact on the transfer of error from initial to final classification phases.

Suitable risk data and instruments that would facilitate the testing of all proposed research questions and hypotheses in the current study are not readily available. This is because, in order to explore the different facets of the proposed inquiries, specific situations are requisite- and these particular situations may be easier tailored into a fabricated data than to be found in the real world. Thus, relying on both conceptual data and actual risk data, random and systematic error are simulated and injected into each

risk instrument to gain insight onto how unreliability and invalidity statistically impact classification.

The risk data are engineered using Monte Carlo Simulation: construction methods making use of random draws from an error distribution and multiple replications over a set of known parameters. This methodology is particularly relevant in situations where the only analytical findings involve asymptotic, large-sample results. Monte Carlo Simulations enables the construction of multiple datasets in a "laboratory setting" that would simulate data in the real world. This allows evaluations concerning the impact of different risk properties on the transfer of error to be made.

For the current study, two main questions are asked: 1) what is the impact of error in risk data on overall classification outcomes; and 2) how does such error impact validity. The study found that risk tools generally have a low tolerance for error. The injection of 10 percent error into risk assessment information produced 25 to 40 percent error in classification outcomes. However, the injection of random error only minimally reduces classification validity by causing the subgroup recidivism/base rates for each category to mildly shrink towards the mean. Different risk tools and factors play a critical role in determining an instrument's sensitivity to error. Specific risk properties such as dichotomous risk items, having fewer risk categories, risk items with lower weights, and having more risk items reduce the sensitivity of error in risk tools. A risk tool's tolerance for error is, thereby, controlled by a confluence of factors.

This dissertation facilitates a better understanding of the interplay between error in risk information and error in classification outcomes. The findings improve

knowledge of the sensitivity of error in offender risk classification instruments. Furthermore, it explains how the sensitivity of error is aggravated or mitigated by the inclusion of different common risk device properties.

Acknowledgements

I would first like to thank my dissertation chair, Dr. Todd Clear, for his insights and guidance throughout this entire process. Second, I would like to thank Christopher Baird and his team for sharing their risk data on which the tests were conducted. Additionally, I am indebted to Dr. Robert Apel, Dr. Joel Miller, and Eric Leneskie for their helpful comments on drafts. Finally, the completion of this endeavor would not have been possible without the support of my family, thank you.

Table of Contents

Abstract……... ii
Acknowledgements……. iv
Table of Contents…… v
List of Tables…….. vii
Chapter I-Introduction……. 1
Importance……. 2
Purpose……... 5
Classification for Risk…… 6
Functions and Classification Characteristics……. 6
Classification and Prediction……. 11
Accuracy/Prediction Issue…….. 15
Base-rates…… 20
Selection Ratios…….. 23
Cutoff Scores…….. 23
Validity in Context…….. 26
Constructing Classification Devices……... 27
Chapter II-Theoretical Framework……... 31
Data Problems……. 31
Data and Omission…….. 36
Transporting Risk Devices……... 37
Error in Application…… 39
Problems Focused in Study…… 43
Chapter III-Standard for Construction and Evaluation…….. 45
Informal Construction Process…… 46
Measures of Validity…….. 50
Evolution and Validity…… 57
Reliability…… 71
Equity…….. 73
Cost and Efficiency Chicago's Public..... 75
Chapter IV-Methodology……... 78
Statement of the Problem…… 78
Datasets……... 79
Procedure for Injecting Error…….. 97
Hypotheses…….. 98
Analytical Plan…… 105
Measurement/Statistical Analysis……..… 110
Chapter V-Assessing the sensitivity of error in risk devices- Analyses and Results.. 113
Results- Hypothesis #1…… 114
Hypothesis #2……... 119
Hypothesis #3……... 122
Hypothesis #4……. 126
Hypothesis #5……... 132
Hypothesis #6…….. 138
Hypothesis #7……... 146
Hypothesis #8…….. 155

Hypothesis #9…… 158
Hypothesis #10…….. 161
Hypothesis #11…….. 165
Hypothesis #12…….. 170
Summary……. 173
Hypothesis Summary…….. 176
Chapter VI-Discussions and Conclusions……... 178
Overview of Research Findings…….. 178
Limitations…….. 184
Policy/ Future Research Implications……. 188
References……... 191
Curriculum Vitae……. 200

List of Tables

Table 1…….. 18
Table 2…….. 19
Table 3…….. 54
Table 4…….. 56
Table 5…….. 83
Table 6.……. 84
Table 7……. 87
Table 8…….. 88
Table 9…….……. 91
Table 10…….……. 92
Table 11…….……. 94
Table 12…….……. 96
Table 1A…….……. 118
Table 1B…… 118
Table 2A…… 120
Table 2B…… 121
Table 3A…… 126
Table 4A…… 129
Table 5A…….……. 135
Table 5B…….……. 135
Table 5C…….. 136
Table 5D…….. 136
Table 5E…….. 137
Table 5F…….. 137
Table 6A…….. 140
Table 6B…….. 140
Table 6C……. 144
Table 7A……. 149
Table 7B…….. 150
Table 7C…….. 151
Table 7D…….. 151
Table 7E…….. 152
Table 7F…….. 153
Table 8A……. 156
Table 11A…… 167
Table 11B…… 168

Chapter 1

Introduction

In biology, classification refers to the arrangement of living organisms into different groups due to their similarities and differences. Separating organisms into different classes or families helps scientists understand what different species have in common with one another. The classification of offenders is similar. In criminal justice, offender classification refers to the disaggregating of offenders into groups of individuals with similar attributes. Classification helps criminologists divide offenders into meaningful groups to serve a specific purpose, which will depend on the goals of agencies (Champion, 1994). Based on a scientific model, it allows offenders to be treated as members of groups for which there is an experience base (Clear, 1988). The actions of other members of the group to which they belong form this experience. Without the ability to group offenders, science in the use of classification is of no help to practitioners in structuring decisions (Clear, 1988).

Misclassifications or the incorrect placement of individuals can occur if sufficient error enters to distort classification. However, initial error does not necessarily translate into final misclassification- this will greatly depend on the sensitivity of the classification instrument to error. In other words, the initial quantity of error may or may not be commensurate of the amount of error that is experienced in the final classification phase. Many factors such as base rate, distribution of data, numbers of classification categories, and number of classification items are speculated to have an impact on the sensitivity of

classification instruments. Hence, a primary purpose of the dissertation is to model the differential effects of such factors on the sensitivity of error in classification devices.

The next chapter will discuss the potential sources of error that may disrupt the function of classification outcomes. There are generally two sources of error: validity and reliability. Validity involves the low correlation found between measures and events, while reliability involves the consistency to which information is accurately reported (Gabrill & Shlonsky, 2000). In particular, error issues are comprised of inaccuracy of offender information, staff bias, definitional dilemmas, inability to identify risk factors, etc. These errors will cause the increase or decrease in an individual's risk score. However, the change in risk score, consequent of the error's initial impact, may or may not necessarily affect the individual's final placement in a risk category. The relationship between initial error and final impact on classification outcome is called sensitivity.

Importance

With the growing amount of research on high-rate chronic offending, many states began to develop and use classification systems with some degree of regularity (Champion, 1994). The idea that a sizeable portion of criminal activity can be attributed to a small portion of offenders encouraged the proliferation of preventative measures and early identification (Jennings, 2006). Research has also shown that the increased monitoring and surveillance of high-risk offenders by law enforcement can increase rule compliance by enhancing the certainty of punishment (Hagenbucher, 2003). Fueled by this reality and the paucity of resources, classification quickly grew to become a cornerstone in corrections and the criminal justice system.

Classification is of fundamental importance in almost every aspect of the criminal justice system. Whether it is conscious (manifest) or unconscious (latent), criminal justice practitioners constantly form decisions about individuals using classification (Champion, 1994). In a technical sense, local law enforcement heavily relies on classification that is based on some preconceived notions about what combination of attributes offenders typically have. The decision to arrest or to initiate investigation during a routine patrol is fueled by some calculation of risk that classifies an individual or a particular situation. And in a much more formal sense, courts and practitioners regularly use classification instruments to inform decisions about: 1) whether to set a bail bond, release offender based on recognizance, or deny b; 2) prosecutorial decision making; 3) inmate management; 4) periodic reassessments of inmates to understand the change in dangerousness; and 5) early release decision or parole (Champion, 1994). Thus, offender classification is a cornerstone to our criminal justice system that which ensures efficiency in every step of criminal justice decision-making.

Due to the paucity in resources, the use of offender classification is necessary for an efficient criminal justice system. The need to objectively identify future risk among offenders emerged from the incapacity of the criminal justice system to adequately provide correctional supervision to the entire criminal population (Champion, 1994; Lowenkamp et al., 2001). The criminal justice system supervises over 7 million offenders each year (Glaze & Bonzcar, 2006). However, notwithstanding the system's ability to accommodate prison spaces for such a large population, there aren't enough prison cells. As a result, the individual differences across offenders and their individual

propensity for future offending make it imprudent to take a "one-size-fits-all" approach to correctional treatment.

Furthermore, assigning a uniform punishment to different offenders irrespective of crime severity or individual risk is against the principle of justice and fairness (Clear, 1988). Thus, the criminal justice system has a daunting task of identifying the risk of every offender in order to determine the appropriate correctional supervision each offender needs. And perhaps more importantly, individuals will benefit little or even be harmed when program treatment is a poor fit, such as when "low risk" offenders are sent to boot camp. In meta-analytic studies of the effectiveness of various crime policies, Cullen and Gendreau (2000), Sherman et al. (1997), and MacKenzie (2000, 2006) consistently found that mistakenly placing low risk individuals into

high-risk treatment programs will increase propensity for offending. If one of the primary goals of offender classification is to reduce future offending, such mismatch in supervision will have the opposite effect on crime. Therefore, agencies greatly benefit from using classification systems because they allow resources and staff hours to be allocated more effectively and optimally (Holsinger et al., 2003; Clear & Gallagher, 1985).

Purpose

Offender classification has many functions and purposes; such purposes determine how the individuals are divided. Thus, the criteria used to divide the groups will vary much depending on the context in which the classification device is used. Champion (1994) describes twelve specific functions of risk classification used in criminal justice, which include: program placement of sentenced offenders, identify offender needs for specialized treatment, early release or selective incapacitation, to determine initial custody level or type of supervision, to evaluate behavioral change in inmates, and to evaluate the need for more prison construction. Thus, based on the specific goals and functions set forth by an agency, classification instruments will contain different criteria for separating groups. For instance, offender classification systems contain items related to risk if their purpose is to place individuals into different levels of custody. If the purpose is rehabilitative, the classification instrument will include needs items that could sort individuals based on their needs.

Classification for Risk

One of the most common forms of classification, especially in correctional contexts, is to group people based on "risk." Offender risk classification is based on risk measurement, but the purpose of risk classification is not to predict individual behavior but to use prediction devices to sort people into risk groups. The distinction is key.

Many definitions exist for the term offender risk classification, but generally, classification identifies individuals for grouping. In attempting to define the term, some scholars have placed emphasis on the procedural aspect of classification. Gottfredson (1987) stated that classification refers to the separation of groups based on some system or rules that have already been determined, and similarly, Champion (1994) defines classification as the procedure of grouping persons based on their specific characteristics. Yet others have embraced a definition that emphasizes more on the purpose and goals of risk classification within the correctional system. Sechrest (1987) argues that the definition of risk classification is often vague about the goals of classification. Concomitantly, Andrews and Bonta (2003) and Brenan (1987) argue that classification is a means of maintaining institutional safety and a means to effectively implement correctional interventions. Despite the variety and subtlety in the interpretations of the term offender risk classification, there is no universal definition.

Function and Classification Characteristics

The properties of a classification device vary depending on the purpose set forth by the agency using it. The designer of the instrument will attempt to "tailor in" specific properties to regulate or manipulate stipulations for each group, subgroup base

rate/selection ratio, group sizes, number of groups, number of classification items, etc., depending on the context and goals. This section explains how the manipulation of these properties is directly related to the intended purpose of the agency. It is equally important to understand how these properties affect the utility of classification instruments and the sensitivity of error.

Classification devices, depending on their goals and purposes, contain different discerning attributes for dividing offenders into subgroups, often called "factors." Such discerning attributes or factors that form the points of comparison and contrast among the groups are not randomly chosen. Instead, the factors used to sort offenders into groups are linked to a purpose, dependent of the context in which it is used. The group characteristics can tell us why and how individuals are classified. As an obvious example, the age of an offender will constitute the discerning attribute used to sort individuals if the classification system seeks to create age groups (i.e. juveniles, adult, elderly) for the purpose of differentially allocating resources based on the unique needs of each age group.

In criminal justice, the primary purpose of offender risk classification is to classify offenders so that varying level of supervision or program intensity can be effectively and appropriately assigned (Holsinger et al., 2003). Offenders are divided into groups based on risk or the propensity for future reoffending. The attributes used to form the criteria for these groups are characteristics found to have some correlation to risk, called "risk factors." Thus, a classification instrument used in a criminal justice situation is usually comprised of variables that are statistically related to a criterion such as re-arrest (Glaser, 1987). More recent classification systems have evolved to include

other sets of discerning attributes, such as need and responsivity (Andrews and Bonta, 2006). They are meant to group people according to need for supervision, one aspect of which is risk. Such classification devices are designed to fulfill a wider variety of specified purposes than mere risk grouping. Despite this theoretical departure from earlier risk devices to include other purposes in offender classification, the basic purpose remains the same; offender classification devices attempt to classify and divide large groups of offenders based on some prearranged set of discerning attributes, whether it is risk, need, or responsivity.

It is common, especially with risk instruments, to use a combination of risk factors that are added up into a single risk score. Each person being classified gets a risk score. The risk grouping is determined by "cut-off" scores. Thus, a hypothetical population's risk scores may range from 0-50; cut-offs may create the following risk groups: 0-12= Low, 13-30= Moderate, 31-50= High.

The population will have a "risk" base rate, which is the frequency of the occurrence in a population (Gottfredson & Moriarty, 2006). For example, if it is being assessed for risk of a new arrest, and 40% of the group experiences a new arrest, the base rate is 40%. When grouped by risk, each group will have a subgroup base rate. For example, those scoring "low" on risk might have a base rate of 20%, for experiencing a new arrest, "moderate" 40%, and "high" 60%.

The determination of cutoff scores and subgroup base rates is contingent on the purpose. After identifying a criterion such as parole violation, re-arrest, or reconviction, the base rate for the criterion is ascertained. The base rate will then aid the device

designer in dividing the group. Taking into consideration of the totality of the circumstances, the designer will then draw artificial lines or cutoff scores, which set the stipulations for each classification category. Typically, the cut-offs will be chosen to ensure that the sub-base rates for each subgroup are vastly and meaningfully different so that different program treatment can be assigned. The overall base rate, thus, has a strong influence on how the categories will be created and how the offenders will be divided.

However, the base rate is not the sole determinant of cutoffs as many other factors are considered. For instance, if the cutoff scores in the example above create a deep positive skew, where most individuals are pushed into the high risk category, this could be problematic if there are not enough program treatment slots for the high risk group. Here, the group sizes created by the cutoff scores do not match the program capacity. To alter the cutoff scores in a way that maximizes the efficiency of treatment of individuals in this agency would mean that less weight should be given to the actual subgroup base rates so that more focus can be given to group sizes. Thus, in this situation, the designer would have to consider the availability of resources in the agency, in addition to the base rates, before drawing the cutoff scores.

Classification device designers also have to consider the number of categories to be created. It is not uncommon to see classification devices divide individuals into five or more categories (Baird, 2009). The question is whether it is necessary and whether it is circumstantially beneficial. If more effective therapeutic program treatment options are available, then the formation of more groups is warranted. However, Baird (2009) cautions that most agencies that utilize classification devices to classify offenders into a large variety of categories do not necessarily have enough treatment options that

correspond with every individual group. Thus, the number of groups a classification device seeks to create is dependent on the resources available in the agency.

The number of factors in a risk instrument is another classification property that is flexible. There is some disagreement about the added utility of incorporating multiple interconnected risk items into risk classification devices. Historically, the earlier classification devices were concise, typically consisting of fewer than a dozen factors (Baird, 2009). However, recent classification instruments have been comprised of many more items. The LSI-R (Andrews & Bonta, 1995) and YASI (Orbis Partners, 2008), for example, contain 54 factors and 117 factors respectively. The multitude of factors in recent instruments represents a sharp departure from the content, format, and goals of earlier instruments.

The creators of the LSI-R justified the inclusion of many more variables by arguing that classification devices should be competent in "risk reduction" (Andrews & Bonta, 1995). Traditionally, risk assessment instruments relied solely on static (i.e., historical, unchangeable) factors such as criminal history, which can be useful for classification purposes but were constrained by an inability to contribute to the effective treatment planning and ongoing evaluation of offenders (Bonta and Andrews, 2007). More recent risk assessments (e.g., Psychopathy Checklist-Revised or PCL-R, Violent Risk Appraisal Guide or VRAG, Self-Appraisal Questionnaire or SAQ; Loza & Loza-Fanous, 2001) are risk/needs assessments that include both dynamic risk factors (e.g., criminal attitudes and companions) and static factors. Dynamic questions assess offenders' current needs (e.g., present employment, criminal friends, family relationships, etc.) to help decision makers gain insights on the offenders' current and ever changing

situation, which is particularly useful in guiding the delivery of rehabilitation service, and is specifically aligned with the goals of "risk reduction" (Andrews & Robinson, 1984; Motiuk, Bonta & Andrews, 1990).

However, there is sufficient evidence available to suggest that relatively brief risk indices outperform longer models (Wagner, 2008). Similarly, Austin, Coleman, Peyton, and Johnson (2003) found that relatively few of the LSI-R factors are significantly correlated with outcomes. When comparing the 42-factor LC/CMI with the 11-factor risk assessment instrument used in Nevada (Onifade, Davidson, Campbell, Turke, Malinowski, & Turner, 2008), Baird (2009) found that shorter devices were able to sort individuals into groups with larger differences in subgroup base rates and better divided individuals into more useful and meaningful groups by significantly shifting individuals away from high risk level. Even studies done by proponents of the LSI-R models frequently determined that most risk factors demonstrate little or no relationship to recidivism (Flores, Travis, & Latessa, 2004). Therefore, it is evident that the purposes for the use of classification instruments affect the number of factors that are used. For the LSI-R, the inclusion of "risk reduction" into the goals of classification systems increases the number of factors, even when its ability to effectively classify individuals can be undermined.

Classification and Prediction

Prediction research in the field of criminology is primarily conducted to aid criminal justice practitioners in the use of risk classification systems (Jennings, 2006). By understanding the "future state of behavior" (Gottfredson, 1987, p.2), prediction

research provides parole boards and correctional staff with the likelihood that an offender will re-offend, as well as supplying information about the likelihood that the offender will abscond or jump bail (Farrington & Tarling, 1985). Therefore, risk classification is constructed based upon research and principles in prediction (Jennings, 2006). When making a prediction, there are four possible outcomes. They are: 1) true positives or cases that are correctly predicted to succeed, 2) true negatives or cases that are correctly predicted to fail, 3) false positives or cases predicted to succeed but do not, and 4) false negatives or cases predicted to fail, but do not. Together, these four categories comprise all the possible prediction outcomes.

Traditionally, validity in offender risk assessment was measured by the degree to which "predictions" about individuals are correctly made. The proportionality of true positives and true negatives against false positives and false negatives is measured to evaluate the validity in "predictions" (Ruscio, 1998). When more cases fall into the true positive/true negative categories, the instrument is deemed more valid.

On the other hand," classification" is not "prediction." Rather, classification is grouping of people based upon risk. In a high risk group, it is known that some people will not fail. But this group is also known to have a higher rate of failure than any other risk groups. Thus, it gives a clear indication that certain individuals need more attention and services, because cases in this designation are inclined to "fail" at higher rates than individuals in other categories (Baird & Wagner, 2000). Therefore, while "classification" seeks to divide individuals, the "prediction" aims to forecast human behavior with precision (Baird & Wagner, 2000).

It is fundamentally important to understand that the dissimilarities between "prediction" and "classification" should not overshadow their shared commonalities. "Risk Classification" is supported and built upon prediction research and methods (Clear, 1988); that is, classification variables or factors are carefully selected based on research in risk prediction. However, the process of dividing offenders has grown to be known as "classification" rather than "prediction" in part because of our inability to accurately identify individuals who will reoffend. Because "risk prediction" intimates a steadfast ability to see into the future and because human behavior is inexplicable, unpredictable, and complicated, Clear (1988) cautions us not to map rules onto the whole range of human behavior in some hard and fast manner, such as in "prediction". The science is not wrong, but the quality of science of human behavior is so limited that human behavior remains too much a mystery (Clear, 1988). Thus, the concepts underlying classification represent a sharp departure from the traditional sense of prediction (Baird & Wagner, 2000).

Risk instruments are not precise, and their ability to identify individuals who will fail is somewhat weak. Validity can be measured by the "hit rate" or the correct identification of failure (true positive) and non-failures (true negatives) (Ruscio, 1998). For example, Schlager (2005) who validated the LSI-R in New Jersey found that the accuracy of correctly identifying true positives and true negatives was slightly better than pure chance. Also, Zhang et al. (2007) used the Area under the ROC Curve to assess predictive validity of risk instruments. AUC varies between .50 (pure chance) and 1.00 (perfect prediction). AUC less than .60 is considered weak, .70 moderate, .80 strong (Tape, 2003). Zhang et al. (2007) confirmed that most validated risk models are only

weakly to moderately accurate (.60-.70 AUC). Similarly, Gendreau et al. (1996) conducted a meta-analysis using Pearson r to measure effect sizes in popular risk selection devices and found the following mean r: .35 (LSI-R), .29 (SFS), .27 (.08), .28 (PCL), and .16 (MMPI Based). This shows that these instruments provide only weak to moderate predictors of recidivism. Therefore, objective offender risk assessments, though deemed much more accurate and predictive than clinical judgment, are only weak to moderate predictors of failure (Gottfredson, 1987).

The problem of accuracy in prediction is even more pronounced for rare events. This is the case because when events are rare in the population, even those who are of high risk will be unlikely to experience the event. For example, if a base rate is 5%, those of high risk of that event will still be unlikely to experience it. Thus, there will be a great deal of false-positive prediction.

"Offender risk classification" is designed to help circumvent the accuracy issue posed by predicting rare events in individuals (Baird, 2009). Risk classification argues that the primary purpose of risk devices is to assign individuals to different groups based on some identified characteristics so that recommendations for proper program treatment can be made. As long as the risk device can meaningfully form different groups, it will have served its intended function. Based on this definition of risk classification, the importance of accurately identifying individuals who will recidivate is downplayed dramatically. When a prediction is wrong, such as when an offender is predicted to fail and does not (false positive) or when an offender is predicted to not reoffend and does (false negative), the instrument is considered flawed. But because classification does not

predict individuals but rather creates groups, the prediction problem is considered less severe.

A risk device can aptly classify individuals into groups, but it cannot fulfill the responsibility of predicting risk. It is a common mistake to think that risk classification devices predict individuals. They do not. They merely create groups of people with different risk rates. Together, the function, utility, and purpose of offender risk assessment instruments have been misunderstood. The distinction between classification and prediction is important for the dissertation because the measurement of validity for risk devices in each context is vastly different.

Accuracy/Prediction Issue

In risk assessment, error is unavoidable because it is naturally a part of risk instruments. From an econometric point of view, disturbance can come from three sources: from the *omission of the influence of innumerable chance events*, *measurement error*, and *human indeterminancy* (Kennedy, 2008). The *omission of the influence of innumerable chance events* refers to the inability to include in a causal explanation the net influence of a large number of small and independent causes. *Measurement error* refers to the inability to accurately measure a variable that is being explained, either because of data collection difficulties or because it is inherently unmeasurable. Finally, *human indeterminancy* refers to the belief that human behavior is such that actions taken under identical circumstances will differ in a random way. A meticulous researcher may be able to minimize error derived from the first two sources, but when it comes to randomness, it is unpreventable. Randomness refers to the notion that the same causes do

not always yield the same effect in the real world, and across different settings. That is, if different people are placed in the same situation time and time again, the reactions, responses, and outcomes will vary due to randomness. Thus, error is inevitably intertwined with any risk device (Gottfredson, 1987). This section explains the prediction model and its inherent problems as an introduction to other problems that are relevant to risk classification. It is, however, important to remain cognizant of the distinction between classification and prediction.

From "sensitivity" and "specificity" perspectives, error that could undermine the integrity of a prediction study has two general effects: false positives and false negatives (Gambrill & Shlonsky, 2000). In criminal justice these two types of error are commonly known as 1) risks to the individual and 2) risk to the community (Wilkins, 1985).

False negatives (FNs) result in costs to the victims of crime and the community. By wrongfully categorizing an as a non-recidivist, this error would come at the cost of having more victimization. Not only are victims directly impacted, but the community as a whole has to pay for false negative errors. Clear (1988) explains, as communities experience more crime, the quality of life is diminished, neighborhoods become less wholesome, and people begin to fear each other. Because of the political and social costs of false negatives, many prediction instruments would rather minimize their occurrence, even though there will be tradeoffs or consequences that directly result in the increase in false positives.

False positives (FPs) are equally deleterious to the fabric of society- their cost affects other citizens. Erroneously categorizing someone as a recidivist forces

unwarranted levels of control. This, in part, is what contributed to the spike in prison population in the 1990s when legislature passed laws that applied a higher level of risk assignment to convicted drug dealers (Sherman, 1997). The social and economic costs are great. Tax-payers must share the burden of paying for expensive correctional control (VanVoorhis & Brown, 1996). Moreover, unnecessary controls are a social cost to the individuals whose freedom is being taken. More recent research indicates that a mismatch between an individual's risk and level of program treatment can also have negative effects. For example, many meta-analytic studies on the effectiveness of different crime policies found that assigning first time offenders to boot camp increases future offending (MacKenzie, 2000; Sherman, L. 1997).

Clearly there should be every attempt to minimize the proportions of both false positives and false negative. But no matter how small the percentage of risk of incorrect classification, there will always be an issue of trade-off of false positives against false negatives. For example, if we over insure against recidivism, we generally tolerate more false positive error and the community will be saved some crime. However, this comes at the price of increased risk of individual offending. If, on the other hand, we are more willing to tolerate crime in the community, there would be more risk to individual freedom. Therefore, the admittance of error is inescapable- the researcher must deal with error and decide on an acceptable level of FNs and FPs (Clear, 1988).

The relationship between false negatives and false positives is complex. In Statistical Prediction in Corrections, Clear (1988) puts these concepts in perspective for us. Drawing from a hypothetical scenario where 1,000 offenders are subjected to a prediction of future risk in felony arrests, Clear (1988) estimates the distribution of error.

Based on a 20 percent base rate, meaning that absent intervention, there will be 200 offenders rearrested for a felony. The following figure illustrates this distribution.

Table 1: Prediction Outcomes

	Actual Outcome			
		Failure	Success	Total
Predicted Outcome	Failure	(A) 100	(B) 100	200
	Success	(C) 100	(D) 700	800
	Total	200	800	

If we assume that our prediction instrument can accurately identify one-half of failures in the group, we will have 200 erroneous predictions divided evenly between false positives and false negatives. According to Gottfredson (1987), a 50 percent true positive rate is considered good for most prediction devices. Within a criminal justice context, the judge would have erroneously incarcerated half of the 200 offenders with a predicted outcome of failure. Moreover, another 100 offenders would have been erroneously released. Thus, the inherent problem with risk devices is that error is unavoidable.

Both types of error are not weighed the same. One difference between false negative and false positives is that the former is more visible to the public. The public is especially sensitive to false negatives because these are the types of error they can see and fear the most. On the other hand, false positives, which result in erroneous incarcerations, are a severe intrusion into someone's life, but because the public does not see the prevalence of it, they don't appreciate the magnitude of the problem. Thus, there is constant pressure on the criminal justice system to over-predict risk in offenders or to reduce false negatives at the cost of increasing false positives. This is the problem that

confronts risk device designers- they must balance out the two types of error while attempting to assuage the public's sentiments towards false negatives because of their lack of objective knowledge about crime.

The level of error can be altered to fulfill a particular goal, but it also comes with a high price tag. For example, if we were to reduce false negatives by 50 percent in the above scenario, it can be accomplished, but it would increase false positives by 450 percent. The error distribution is drastically altered by the mildest attempt to reduce false negatives by 50 percent. Stated differently, category (B) has now increased by 450 percent. The figure below illustrates the problem. The numbers alone make the problem seem mild and acceptable, but if we put them into context, the criminal justice system will have to infringe upon the freedom of that many more people in order to insure a modicum of reduction in false negatives. Whether a 50 percent decrease in false negatives outweighs the cost of increasing false positives by 450 percent is a separate philosophical debate. Clear (1988) shows that it is statistically imprudent.

Table 2: Prediction Outcomes (Continued*)*

	Actual Outcome		
	Failure	Success	Total
Predicted Outcome			
Failure	(A) 150	(B) 450	600
Success	(C) 50	(D) 350	400
Total	200	800	

Thus, practitioners must determine the extent and type of error they are willing to tolerate, given the instrument's proclivity for false positive and false negative error. The frequency of each type of error will dictate the amount and extent to which funding is allocated for programming and supervision. Determining the proportions of the two

types of risk is paramount to ensuring that proper resources are allocated in keeping with the goals set forth by the criminal justice system.

Validity within the context of risk prediction is typically measured with predictive validity tests. As mentioned earlier, the true positive rates that are identified by such validity tests give an incomplete description of the problem of validity. As important is the investigation of the true negative, false positives and false negatives that make up all the prediction outcomes. In an ideal situation where an instrument is perfectly valid, there are zero cases for false negatives and false positives. The increase in cases in these two categories will signify a gradual increase of invalidity in the instrument. Though the prediction model and its related problems discussed here are not directly related to classification, it sets the stage for our understanding of other problems relevant to risk classification, such as the problem of base rate.

Base Rates

The base rate for any given event is the frequency of the occurrence in a population, which is usually expressed in percentages or proportions (Gottfredson & Moriarty, 2006). For instance, if we were to create a prediction device for parole revocation, the criterion would be parole revocation. The base rate would be the number of failures compared to the entire parole population. There are two major problems related to the base rate problem.

First, most statistical methods for measuring predictive accuracy in risk devices are sensitive to the base rate (Gottfredson & Moriarty, 2006). This is the true offending rate for a specific criterion within the entire population. Clear (1988) explains that

knowing the subgroup base rate is important because it lets us know "how high is high?" (p.21). An instrument's ability to accurately classify is of limited utility unless we know how often the event of interest (base rate) occurs for the subgroups. He suggests that most selection devices are constructed independent of the actual base rate.

Gottfredson and Moriarty (2006) explain that knowing the true base rate is not possible. In predictive analyses, researchers attempt to construct, validate, and assess the accuracy of risk instruments predetermined selections. In other words, risk assessment instruments depend largely on actual offending statistics to find appropriate cutoff scores, selection ratios, etc. The small differences between actual offending and the known offending rates could cause large margins of error in the instrument. The current study acknowledges that there may be great discrepancies between such offending rates. This notion is a primary justification for the study.

The problem is that it is difficult to assess the true base rate for an event such as reoffending because the only offending rates that can be quantified are the ones that are known and visible. Unreported crime itself produces a large discrepancy between the known base rate and the true base rate, thereby compromising accuracy in predictions.

Second, the difficulty of predicting outcomes increases as the base rate diverges from .5 (Meehl, 1954). That is, as the base rate increases or decreases from 50 percent, there tends to be more prediction error. Borrowing from Gottfredson's (1987) explanation of the problem, suppose that the base rate of parole failure is 20 percent. Given this information, if we make a risk prediction that no one will fail, we will be correct 80 percent of the time. We will also be wrong 20 percent of the time- there is no

way of telling which 20 percent will fail. Thus, by simply over-classifying all the offenders as non-failures, the error will be 20 percent, based on a 20 percent base rate.

Now, if we assume that a strong prediction device was developed and has the ability to accurately predict parole revocation with 78% accuracy, which is far higher than the accuracy of most classification instruments, the issue becomes clearer. Despite the superior accuracy in this hypothetical risk device, we would still be better off in expecting that no one will fail on parole if the base rate for the outcome is low. In fact, there would be far less error in this case, if a risk device wasn't used at all (Gottfredson, 1987). If the base rate is 5 percent, the blanket assumption that no offender will fail will be correct 95 percent of the time. The predictive accuracy of not using any instrument will be 17% higher than the hypothetical risk device. Thus, the base rate is an important mathematical component to consider, especially when the predicted events occur infrequently (Meehl, 1954). Nonetheless, most contemporary research reporting neglects the base rate factor (Gottfredson & Moriarty, 2006).

The base rate is an important component to risk prediction, but it also has important implications for risk classification. Jenning (2006) says that the benefits from the expansion of prediction research directly enhance the effectiveness of classification devices because risk classification instruments are developed based on research and principles in prediction. Thus, increasing our understanding of prediction problems will directly benefit the effectiveness of risk classification.

Selection Ratios

The selection ratio is the proportion of people in a given population who are predicted to participate in an event of interest (Gottfredson, 1987). If the criterion is parole revocation, for instance, the people who are identified to fail will make up the selection ratio. The error associated with selection ratios is dependent on the base rate. Farrington and Tarling (1985) argue that the greater the discrepancy between selection ratios and base rates, the greater the problem with error. Put differently, if a base rate for a criterion in the entire population is 5 percent and the selection ratio for a specific group is 50 percent, there would be higher levels of error than if both of these percentages were roughly equal. This is similar to the low base rate problem that was discussed earlier. The divergence between these two ratios constitutes the problem associated with low base rates.

This becomes a problem because more false positives will occur when the selection ratio is disproportionately higher than the base rate. For example, as more offenders are predicted to belong to a high risk group, the probability that a greater number of offenders will be identified as posing risk when they actually do not will increase. On the other hand, as the base rate increases, it is more likely that someone who was identified to be low risk will pose a risk- false negative.

Cutoff Scores

The cutoff score is important because it determines the selection ratio. In other words, they are superficial cutoff points determined by the instrument designer to serve a particular purpose. This score is used to create risk categories and will influence the

number of false negatives and false positives (Clear, 1988). Social and political goals usually guide instrument designers in setting such thresholds for risk. For example, if prisons are overcrowded, one could manipulate cutoff scores such that more offenders will be classified to low risk categories. Conversely, a "get tough" policy may precipitate change in these scores such that more offenders are identified as high risk.

Ideally, the rate of recidivism for each risk category should be substantially different from each other. In other words, the cutoff scores determined by the device designer should partition the risk level categories so that high risk groups have higher recidivism rates than medium and low risk categories. Though the creation of groups with distinctly different base rates is the commonly accepted goal, there is not a steadfast rule as to what numerical representation in the cutoff scores constitutes validity or invalidity. The general rule for drawing cutoffs is that the subgroup base rates should be sharply different. For example, validation studies in Nevada confirm that the LS/CMI risk instruments is valid when the base rate from the high risk level quadrupled the recidivism rates in the low risk categories, i.e. 9% in low, 24% in medium, and 45% in high risk level (Onifade et al., 2008). Though a strict standard does not exist, cutoff scores that achieve such sharply different recidivism rates in risk levels are considered good.

Invisible lines between risk categories are artificially drawn by selecting cutoff scores. Device designers have much discretion in deciding where cutoffs between risk levels are drawn. Thus, risk devices do not objectively identify risk, but instead, are influenced by a multitude of issues that may or may not have anything to do with actual risk. Different cut-offs could be considered acceptable as long as different risk groups

contain starkly different levels of propensity. However, the instrument would begin to lose its classification merits when cutoff produce sub-base rates that are not different (Van Voorhis & Brown, 1996). To avoid this problem, Clear (1988) encourages systematic evaluations of cutoff scores and propensity for each risk group. As mentioned earlier, budget limits, "get tough" policies, and social pressures can influence the criminal justice system. These forces can also influence risk instrument designers, and thereby undermine the objectivity of cutoffs.

Most risk devices seek to divide individuals into 3 or more levels of risk, but there is little practical utility for systems that divide individuals into, for instance, 8 categories of risk. There are two reasons. First, the levels of risk within a classification system are, sometimes, not aligned with any programmatic or service delivery feature (Baird, 2009). From a practice perspective, Baird (2009) suggests that we ask: what prediction is being made at mid-range risk levels? In fact, employing a risk tool with multiple mid-range categories is disconcerting and irrational when an agency does not have varying levels of supervision to match the different risk groups. Such disparities between an instrument's classification and its actual application create a conundrum. One could, however, make the argument that middle categories are designed to deal with the potentiality of error. The creation of multiple mid-range risk categories could be a ploy to circumvent blame for misclassifications. Without the buffer category, a one point difference may cause an offender's classification to go from high risk to low risk or vice versa, and thereby increasing the instrument's sensitivity to error. Classification systems may be much more sensitive to error once middle classification categories are removed.

Second, it may be more judicious and beneficial for agencies to ensure parsimony and limit the division of individuals into no more than 3 categories of risk. Assuming that proper supervision could be matched to each risk group, the differences between the various levels of supervision may not be meaningfully different enough to justify having so many risk categories (Todd Clear, personal communications, 2013). Beyond the basic requirement that risk classification and supervision should be matched, the type of supervision needs to have a direct link to risk reduction. As such, it is a waste of resources to allocate different levels of supervision if there is no crime reducing impact. Assuming that, for instance, a low risk individual would receive no supervision, a moderate risk individual would receive supervision once a month, and a high risk individual would receive supervision twice a month, the classification system would only be useful if the "twice a month supervision" will have an actual impact on an individual's proclivity for reoffending. In reality, there only exist a handful of pragmatic responses that would effect any positive change. Thus, one needs to ask two basic questions when considering the number of risk categories for a risk device. They are: 1) is there a sufficient amount of varying levels of supervisions to which the different groups can be matched, and 2) Are the varying levels of supervision meaningfully different in their ability to reduce the likelihood of reoffending?

Validity in Context

On the most basic level, validity refers to the extent to which a concept, conclusion, or measurement is well-founded. The threshold for validity, perhaps, is whether a classification device serves the intended purpose. Holsinger et al. (2003) aptly show that many different classification instruments exist, and each is different in terms of

fulfilling certain needs of agencies. It is strongly recommended that the instrument be compatible with the needs and goals of the agencies that use it (Holsinger et al., (2003). For instance, if there are budgetary cutbacks in a department where program treatment for high risk groups are lacking, a classification device that places the majority of offenders into this high risk group would no longer be valid. Instead, the classification device should seek to shift offenders from high risk to lower risk categories (Clear, 1988). This will reduce the constraint caused by the financial insolvency of the agency. Thus, at the fundamental level, the capacity to fulfill an intended responsibility constitutes validity.

Constructing Classification Devices

In order to better understand risk assessments and error, it is critical to understand how such instruments are designed. Risk assessment instruments are created by using one of several standard statistical methods (Brennan, 1987), usually some form of linear regression analysis. And though different statistical methods have been adopted, they appear to have roughly equivalent effectiveness in practice (Gottfredson and Gottfredson, 1979). Similar series of steps are followed in all competent design strategies for constructing risk classification (Clear and Baird, 1987). The following is a summary of these steps.

Steps in Design of risk Assessment Devices

Step 1: Development of a Study Sample**.** Using some type of official record, a representative sample of closed cases is taken to construct the risk instrument. From these cases, the variables that are most commonly associated with the failure criteria (e.g.,

reoffending) are coded (Gottfredson and Gottfredson, 1986). The sample size should be large enough to create reliable estimates.

Step 2: Dividing the Sample. The cases that form the sample are then randomly divided into two groups, a "construction" subsample and a "validation" subsample. The "construction" subsample will be used to develop the prediction model, where as the "validation" subsample is used to test the reliability of its estimates. The validation process allows the designer to see whether the prediction outcomes are a product of "chance correlations", which occur in the construction analysis (Clear, 1988).

Step 3: Constructing the Model. Using some statistical model such as a multiple regression, the variables that were previously selected from the cases are added together. The factors that are sufficiently correlated to the criterion variable are included in the "statistical model", while the other variables are discarded.

Step 4: Validating the Model. Here, the newly formed "statistical model" is tested on the "validation" subsample for the purpose of ensuring validity. This provides another opportunity to see whether the model is sufficiently correlated with the failure criterion.

Step 5: Monitoring and Revalidation**.** Steps 1 through 4 are periodically repeated to ensure that the model is updated to reflect the changes in the offender population profiles. Clear (1988) warns that monitoring and revalidation are critical to determine whether the instrument needs to be updated. However, jurisdictions seldom revalidate these models, which can lead to invalidity.

The process outlined above is a standard procedure for constructing risk assessment instruments. Though its construction seems straightforward, statistical

models are fraught with error and mistakes that are both known and unknown to the model designer- the next chapter will discuss the sources of such error.

It is not yet known how much error is can be tolerated without reducing a risk device's ability to effectively classify. As a result, the literature suggests that a risk device should be frequently revalidated by the agency using it (Van Voorhis & Brown, 1996; Clear, 1988). The general assumption is that when enough time has elapsed, error and misclassification will increase- this is discussed in the next chapter. Despite the many cautionary tales about how classification effectiveness reduces as time elapses, it is not yet fully known how time specifically impacts classification outcomes. But still, there is not a steadfast rule for determining the amount of time that is required to effect enough change/misclassification in a risk device that would render it invalid.

This Study

How validity is conceptualized for each risk device is contingent on its stated goal. For offender risk classification, validity is represented by a combination of elements such as base rates, group sizes, and the capacity to serve an intended purpose. It will be argued that there are known and unknown error issues that create invalidity and misclassifications. Furthermore, these error issues can be intensified or mitigated by the properties in classification devices. How such properties impact the transfer of error into final misclassification and how error can render a risk device invalid are defined as the sensitivity of error.

This dissertation will model the effects of error on offender risk classification devices. Using Monte Carlo Simulations to replicate multiple datasets that represent real

data, different levels of error will be forced into different models with varying properties and characteristics. In each scenario, two general variables that are tested: 1) the level of error necessary to soften the validity of classification devices, and 2) the combination of risk device characteristics that would impact the transfer of error. For each situation, the validity of risk instruments will be measured by the base rates, subgroup base rates, number of cases in groups, and their theoretical connection to intended purposes of classification.

Currently, risk assessment instruments are seldom revalidated. Revalidation is needed when an instrument is transported from one jurisdiction to another; when much time has elapsed since the last revalidation; when the demographics of the offender population have changed significantly; or when a practitioner recommends so, because a high level of misclassification is experienced (Clear, 1988).

By understanding the tolerance of error in these devices, one can begin to understand when and why risk devices should be revalidated. More specifically, practitioners will benefit from the knowledge of knowing how invalidity affects offender risk classification. The need for revalidation will no longer be born out of whim and conjecture. Furthermore, device designers will directly benefit from the knowledge of how different device properties affect the transfer of error.

Chapter 2- Theoretical Framework

This section concerns issues in predicting and classifying offender risk as a justification for the current study. Since risk classification devices are constructed using prediction methods, the problems that plague prediction devices would invariably affect validity in classification. For both risk prediction and risk classification, the different types of error can never be fully circumvented; researchers can only attempt to minimize known errors (Clear, 1988). As was mentioned earlier, error problems generally fall under two categories: invalidity and unreliability. This chapter will take a better look at the various sources of error by which risk assessment information are affected. The distinction between initial and classification error needs to be elucidated. The study argues that all sorts of mistakes and error are intertwined with the information that goes into such risk assessments. On the other hand, classification error refers to the incorrect placement of individuals into a risk designation due to the existence of initial error. The impact of initial error problems may be so enormous that final misclassification of individuals may be disproportionately greater than initial error, or the impact may be so mild that virtually no misclassifications will result- this will depend greatly on the sensitivity of error. To date, no research has been conducted to ascertain this. To better understand the origin of initial errors, this section will review the existent literature on potential sources from which error is born.

Data Problems

Risk devices are constructed typically by using some form of official data (Baird & Wagner, 2000). Thus, the validity of official records is paramount to the validity of

classification devices. This section concerns the more prominent problems within the data from which risk devices are constructed. However, Baird & Wagner (2000) offer comfort in saying that "there is no evidence to suggest that the patterns derived would change significantly if every event was detected, recorded accurately, and responded to in a consistent fashion by all decision makers" (p.847). In other words, given that it is possible to reduce error in the official data, the actual rate of failure in offenders may be higher (or lower) for each known subgroup base rate, but error should proportionally affect all categories. Thus, Baird and Wagner (2000) suggest that "error" bears a proportional relationship to the reported rate for each risk group, thereby duly canceling out the effects of data error when the problem is looked at as a whole. However, the notion that such data problem will cancel-out error remains speculative. This section discusses the measurement of outcomes and related definitional dilemmas.

The outcome for selection designs is often called the criterion. Nearly all prediction studies utilize some form of offending record (e.g., arrest, conviction, incarceration, or parole revocation) as the criterion (Baird, 1991; Farrington & Tarling, 1985). According to a review of 47 studies that explored the predictive validity of the LSI, LSI-R, LSI-OR, LSI-CMI, LSI-R:SR, and YO-LSI, the three most commonly used criterion variables are re-incarceration, re-arrest, and reconviction (Vose et al., 2008). Clear (1988) warns that many common criteria are statistically only marginally related to each other. In addition, Van Voorhis and Brown (1996) aptly says that the importance of a criterion is commonly overlooked, but the criterion has specific effects on the base rate and can have a bearing on the types of predictors used. Thus, a selection design using a specific criterion can only be used to predict that specific outcome.

Generalizing outcomes onto dissimilar circumstances can threaten the external validity of a risk device. In risk prediction studies, it can happen in one of two ways: 1) methodological issues in collecting criterion variable information and 2) over-generalization of findings.

First, there are real issues in finding a proxy that most represents actual offending. As was mentioned, the purpose of any offender risk classification is so that a person's risk of offending can be evaluated. However, the common criterion variables (e.g. arrest, conviction, and parole revocation) are only crude estimations of offending (Baird, 1991). Not all criminals are apprehended and officially processed into the criminal justice system. For most crimes known to the police, nobody gets arrested, let alone getting convicted. Thus, all existing official records of offending are a gross under-estimation of the problem. For instance, of some 3.6 million household burglaries in the United States in 1999 (Felson, 2002), only 200,000 arrests resulted. Only about 2% of burglaries lead to a conviction, and fewer still to incarceration (Felson, 2002). For more common crimes, such as drug offenses, the chance of being punished is much smaller. Thus, there is weak connection between what the data used express (i.e. police efficiency) and what the data are supposed to measure (offending). If what Felson (2002) suggests has any merits, then using arrest data as a criterion variable for risk assessment instruments is flawed because it only captures roughly 5% of all offenses. This affects generalizing because so much error and disturbance exists in the criterion variable.

Second, a viable way to quantify true offending rates of individuals does not exist. Researchers must work with what is available to them. Arrest rates, convictions, and parole revocation are the most widely collected, consistent, and useable data the criminal

justice system has on offending (Farrington & Tarling, 1985). The scholarship has found that arrests and convictions represent most actual law-violating behavior- the correlation between them is so high (r=.80) that either is a viable principal outcome measure (Baird, 1991). Maltz (1984) seems to agree that arrest charges are generally more descriptive of offender behavior than other charges levied by the prosecutor. Maltz (1984) argues that arrest data are more useful than prosecutorial, court, and correctional data because the latter are harder to obtain, sometimes unavailable, and generally, less accurate.

Convictions and parole revocations also make good indicators of predictive outcome. Many researchers prefer the use of convictions to parole violations. Of all the studies conducted on the predictive validity of the LSI during 1982-2008, 20 percent used reconviction data and 3 percent used parole violations. Despite the different criterion variables, the majority of validation studies on the LSI conclude that the instrument is a valid predictor of recidivism (Gendreau et al., 1997; Barnoski & Aos, 2003; Simourd, 2004; Holsinger et al., 2003).

However, aside from the obvious disparity between the true offending rate and estimations of it, there are other limitations associated with these outcome variables. Baird (1991) cautions that arrests are only allegations that may have a limited relationship to actual behavior- especially in groups that are frequently profiled, questioned, arrested and scrutinized more, such as minorities and parolees. Similarly, for convictions and parole revocations the limitations stem from their tendency to measure criminal justice behavior. For example, the lack of funding during some economic downturn can increase caseloads for parole officers, which subsequently reduces their ability to effectively follow-up and detect violations in parolees. Even though these

criterion variables are supposed to reflect offending, they are also dependent on criminal justice policies and behavior. Thus, how offending is operationalized and measured affects the external validity of the design, especially if the findings are generalized to other outcomes. Here, the source of external validity threat is, by default, inherent in faulty data collection and faulty operationalization of the criterion.

The second source of external validity threat comes from over-generalizing across different offense types. For example, the LSI-R poorly predicts violence, spousal abuse, sexual violence, etc. In the recent past, the validity of the LSI-R for offenders with specific offense types has been studied. Manchak et al. (2007) expressed concerns about the utility of generic risk assessments for specific populations such as violent offenders, suggesting that more specific tools should be tailored to this population. The traditional risk factors for general recidivism are, at best, loosely correlated with violent recidivism (Manchak et al., 2007). Scholarship exploring this has shown that other risk factors are better in predicting violent offending, such as: nature of the current complaint, childhood abuse of the parent, number of prior complaints, alcoholism of father, impulse control, etc. (Van Voorhis, Cullen, & Applegate, 1995). Correctional institutions can employ instruments with known utility in predicting violence recidivism, such as the Michigan Family risk Assessment of Abuse/Neglect, the Psychopathy Checklist, or the Level of Service Inventory-Ontario Revision. The latter instrument (LSI-OR) has produced acceptable predictive correlations among subgroups of sexual offenders, domestic violence offenders, and offenders with mental health problems (Girard & Wormith, 2004). Thus, risk assessment instruments developed in one interim may not be transferrable to different types of risk.

Finally, outcomes are very important and specific to selection designs. Even more important is how they are operationalized (Gambrill & Shlonsky, 2000). Prediction and classification are made more difficult as a result of vague definitions of outcome measures. In some cases, the outcome measures, such as recidivism, are not sufficiently defined to build accurate prediction models.

The obvious problem with risk assessment is that there is no universal or perfect measurement of offending. Various criterion variables (e.g. arrest, conviction, and parole violation) are commonly used, but they merely represent feeble attempts to estimate the problem of offending. The actual offending rates of individuals, however, are not completely visible to criminal justice agencies and researchers. Thus, there will always be an immeasurable margin of error inherent to selections designs due to the inability in finding a proxy that will reflect on actual offending.

Data and Omission

The effectiveness of offender risk classification can also be significantly undermined by our tendency to omit important variables. In a recent newspaper article from The Atlantic on the research of Richard Berk, a new wave of offender risk classification devices is developing (Labi, 2012). Richard Berk and colleagues (Sherman, Barnes, Kurtz, and Lindsay) identified new predictors to risk that, in combination, are reported to be more accurate and predictive than existing risk devices (Berk et al., 2009). Berk et al. (2009) uses "statistical learning approach that makes no assumptions about how predictors are related to the outcome" (p.1). And despite their orientation, variables of all sorts are mined from large databases. From racially biased to irrelevant factors

such as “shoe size”, every possible factor is explored. As a result, Berk et al. (2009) has uncovered a myriad of variables that are more predictive, and that have been traditionally ignored.

The work of Berk and colleagues (2009) illustrates an inherent problem with research in offender risk classification; that is, relevant variables are often omitted based on superficiality. Normally, predictors are chosen based on certain procedures that are limited by appeal, its logical relevance to the outcome, convenience and cost (Tarling & Farrington, 1985; Loza & Loza-Fanouys, 2001), theoretical loyalty (Baird, 2009), political and ethical rectitude such as the inclusion of race biased factors (Tonry, 1987), ease of staff interpretation (Flores et al., 2004; Lowenkamp et al., 2004), and social acceptance by staff (Haas & Detardo-Bora, 2009), etc. Thus, the effectiveness of classification is sometimes hampered by our inability to account for the net influence of a large number of small and independent causes. The work of Berk and colleagues (2009) sheds light on the tendency of researchers to disregard irrelevant variables in risk classification studies, which Kennedy (2008) argues is one of the three sources of error for research in the social sciences. Thus, the findings are evidence that many variables are systematically overlooked, thereby contributing to error and disturbance in offender risk classification. This is a major limitation of the science behind risk assessments.

Transporting Risk Devices

The transferability of risk screening devices across jurisdictions is problematic and is a potential source of error. For example, a risk classification system that works well in Newark, New Jersey may not be predictive in a smaller city with dissimilar

characteristics and demographics, or vice versa. Though major risk factors such as criminal history and peer associates change little from jurisdiction to jurisdiction, frequent revalidation has many benefits (Holsinger et al., 2003). Transporting risk devices creates validity issues; they can come from two areas: population variance and difference in external attributes. Population variance refers to the collective profile exhibited in a group while difference in external attributes refers to the difference in neighborhood context. The former has already been discussed in the section about data problems.

The overall demographics and characteristics of a population in one jurisdiction may be starkly different from those of another. For example, much research has shown that offenders who return to disadvantaged and downtrodden neighborhoods recidivate at a greater rate than those who return to "resource-rich" or affluent communities, after controlling for individual-level factors (Kubrin & Stewart, 2006). Thus, a risk assessment instrument that works well in one jurisdiction may do poorly on another, depending on the neighborhood context.

The idea that offenders with similar individual risk profiles (based on criminal history, employment and residential history, etc.) are more likely to fail if they return to disadvantaged high risk communities represents a potential challenge when a risk instrument is transported from one jurisdiction to another. Failure seems inevitable, especially when released offenders are returned to the neighborhoods with the highest crime rates in the nation (Travis, 2005). Unfortunately, these are the cities to which most offenders are released (Travis, 2005). Thus, the base rate for reoffending could vary greatly across different locations, and if caution is not exercised (Clear, 1988),

researchers can run the mistake of assessing risk with an instrument that has no validity for a particular setting. Different neighborhoods could also vary in their propensity for crime by offering more opportunities. Felson (2002) has identified a clear nexus between opportunity and crime. For example, there tends to be more property crime and theft in areas where there are viable avenues to dispose of stolen goods via pawn shops, car junk yards, used-appliance stores. Thus, the neighborhood and the immediate environment can be a powerful determining factor for crime. Holsinger et al. (2001) aptly suggest that revalidation research will permit benchmarking or "create risk categories that are germane to specific jurisdictions or correctional strategies" (p.5).

Areas can also differ in their resources to released offenders (Kubrin & Stewart, 2006), social exclusion from invisible punishments (Travis, 2002), political stance towards offending and penalty, police culture, law, and over or under crowding in prison (Petersilia, 1999). Therefore, cross-validation is an important process that needs to be undertaken whenever a risk instrument is borrowed from another jurisdiction (Gottfredson & Moriarty, 2006)- a different base rate will affect cutoff scores, which will compromise the overall effectiveness in classification. Wright, Clear, and Dickson (1984) illustrate that consequences of the wholesale adoption in several jurisdictions of devices developed in one locale can be severe. It is, unfortunately, common practice today (Gottfredson & Moriarty, 2006).

Error in Application

Finally, there are error issues related to the administration of risk devices by staff. Even with perfect validity in risk devices, where measures are directly correlated to

outcomes, error can emerge from unreliability, which involves the consistency to which information is accurately reported (Gabrill & Shlonsky, 2000). Currently, studies on the use of risk assessment have shown that certain risk instruments are less reliable, suggesting that human error during the administration phase is problematic.

Most risk models are designed with efficiency in mind. Cross-referencing an offender's intake profile will typically give all the information necessary to complete the various items (Van Voohris, 1996). For example, the Ohio Department of Rehabilitation and Correction put forward the following requirements for risk classifications to help guide agencies and to facilitate efficiency: 1) the information needed to complete the instrument is consistently and readily available, 2) the variables are easy to be coded by different users, 3) variables are consistently correlated to the outcome, 4) the variables have face validity or seem to be relevant, 5) the instrument is statistically accurate, and 6) the system is efficient to administer (Van Dine, 1993). This simple guideline set forth by the Ohio Department of Rehabilitation and Correction shows the importance of parsimony and simplicity in instruments. Therefore, sophisticated instruments may be more thorough and accurate, but they run the critical risk of overburdening staff and reducing efficiency.

Efficiency during the use of the instrument can be impeded by implementation issues, staff training, interpreting responses, staff attitudes, and administrative overrides. Most actuarial risk devices were first designed to eliminate human decision-making because discretion leads to subjectivity and error. In particular, this is what initiated the departure from clinical assessments to actuarial based assessments (Champion, 1994). Risk assessment gained increased effectiveness when it excluded clinical judgments

because it directly reduced unreliability. However, this method is not perfect, because despite the rigid structure afforded by actuarial assessments, humans are ultimately responsible for the task of administering these instruments- allowing subjectivity to enter. Therefore, reliability issues still exist, but arguably to a lesser extent.

Negative attitudes of staff towards a particular risk instrument may have an unfavorable impact on the implementation process. For example, Haas and DeTardo-Bora (2009) found that a large proportion of correctional staff did not favor the use of the LSI-R. Only fifty percent of case managers and counselors, and 1 out of 17 parole officers were supportive of it. And as a result of this widespread negativity, only 4 out of 10 staff charged with the implementation of the initiative said they had used the results of the LSI-R to develop reentry case plans for their caseloads. The other six out of ten staff followed their own personal assessment when forming decisions about an offender's risk level, which defeats the primary purpose of using actuarial methods. With all things equal, 60% is an unusually high rate of administrative override that warrants inquiry. "Overriding" refers to allowing the human decisionmaker to decide how to weigh contingencies or special circumstances that requires special attention. Clear (1988) suggests that 15-20 percent of administrative overrides are acceptable. If there are very infrequent overrides, the staff may be over-reliant on the instrument to help them form decisions. And if there are too many overrides, it can be taken that staff are not finding the system useful, as is in the case in Haas and DeTardo-Bora's study (2009). It brings into question of whether expensive risk instruments should even be administered when staff routinely dismisses the instrument's recommendations. General distrust of a risk

instrument can further engender indifference in staff, which may increase unreliable classifications in offenders.

Error can also enter when staff training and experience are insufficient. Lowenkamp, Latessa, and Holsinger (2004) found that the number of years of experience in working with a particular risk device was positively related to predictive validity. Non-professionalism in the staff was synonymous with more mistakes and inconsistencies when administering assessments. Similarly, the Washington State Institute of Public Policy (2004) found classification was more effective in reducing recidivism when it was properly implemented. Some assessments are more sophisticated than others, which require skill and experience to administer properly. For instance, the LSI-R item "prosocial values" rely on their own experience and expertise to determine whether an individual is criminally inclined. Thus, training and experience in using a specific instrument is critically important in risk classification.

Interpretation reliability is another area of concern. Reliability refers to the degree to which the placement decisions of offenders will be consistent across different staff. Many recent and more sophisticated risk assessment instruments, such as the LSI-R or COMPAS, use both static and dynamic variables. As previously mentioned, dynamic variables are assessed via interviews with the offender, which requires careful staff interpretation. Austin et al. (2003) found that variables with the highest inter-rater reliability were static variables, such as criminal history and education/employment variables. Conversely, many dynamic variables in the LSI-R were less than 80% reliable, with some variables having less than 60% agreement. Thus, there is a high error rate in placement decisions that comes from unreliability (Gambrill & Sholonsky, 2000). This

illustration goes to show how the tendency for mistakes tends to increase when the task is more difficult, such as when the staff needs to use his/her discretion to discern or evaluate a situation.

Though the findings generally favor less complicated static variables over dynamic variables because of their simplicity and reliability, Van Voorhis and Brown(1996) cautions about this. She argues that simplicity can lead to inaccuracy and reliability issues, especially when the instrument is so quick and efficient that it doesn't require enough thought. Such risk items tend to get answered in a cursory manner because they are deceptively easy, which contributes to the reliability problem. Similarly, Gottfredson and Gottfredson (1980) warn that it is not as easy as it seems because criminal justice records used in the course of completing static variables are difficult because they are generally unreliable. Thus, both static and dynamic risk variables pose difficulties for staff, which can reduce classification accuracy.

Problems Focused in Study

In offender risk classification, error in measuring predictor and outcome variables can come from: 1) unreliable information from official records; 2) staff distrust, non-professionalism, mistakes, and subjectivity; and 3) transporting risk devices to different settings. Each of these potential sources of error can directly impact the independent and/or outcome variables, thereby distorting the risk information from which risk tools are constructed. Thus, the current study will assume that different levels of error are derived from these places.

Conclusion

Error finds its way into offender risk classification, and will continue to impact the effectiveness of classification despite efforts to abate it. The literature has demonstrated that there are multiple sources of error, though it is not completely clear how much error enters into these instruments. Shortcomings in the methods for quantifying error will ineluctably pose a credible obstacle. Therefore, notions about how much error exists remain largely speculative. The primary objective of the dissertation is not to determine the level of error that transpires because error is expected and inevitable. However, the dissertation aims to understand how such error problems affect final misclassification, which is influenced by an instrument's sensitivity to error.

Chapter 3- Standard for Construction and Evaluation

Today, the goodness of a risk assessment instrument can be judged by its validity, reliability, equity and cost-effectiveness (Baird et al., 2012). The construction of such instruments is fundamentally useless if the instrument does not judiciously take into account all of these competing requirements. The best-case scenario, obviously, is to construct an instrument for which each of these requirements is maximized. Unfortunately, in the real world of risk assessment, risk devices are neither built nor applied in a vacuum, and such requirements are often in fierce competition with one another. For example, the validity of a risk assessment instrument would be significantly enhanced if we could avail staff with more time and resources, so that the full circumstances for an individual could be weighed and factored into the determination of his/her risk. Yet despite this reality, there are real pressures to restrict such assessments and limit them to 20 minutes, for example. Striking a healthy balance between these requirements in a risk device is a complex task for which risk device designers are responsible.

The daunting task of building "good" risk assessment instruments is admittedly worsened by the need to balance and consider validity, reliability, equity, and cost-effectiveness. But on a brighter note, such goals help structure, guide, and standardize decision-making during the construction of risk devices. In fact, these established goals help abate the arbitrariness involved in the construction process. Anyone who is familiar with the construction process knows that it involves a variety of decision points for which there is a lacking standard. For instance, the number and type of risk items to be included, different cutoff points, or the number of risk categories that which comprise an

instrument, are some aspects that require much subjective judgment, mainly because there lacks a clear standard for what a risk instrument should look like. As such, these instruments are more an art, than it is a science. In many cases, risk devices are the end result of multiple layers of arbitrary decisions. Thus, these goals and requirements are helpful in guiding some of the decision points, but it only provides mild levels of relief to risk designers. By linking decisions with established goals, some of the arbitrariness of their decisions is reduced by being able to justify and attribute decisions to goals, which would otherwise be considered capricious.

This section seeks to better understand the different standards by which risk devices are constructed and/or evaluated. It further argues that it is not enough, underlying the importance of the current study and its finding on sensitivity. The understanding of how different risk device properties impact the sensitivity of error will hopefully establish another requirement based on which selections and decisions could be made.

Informal Construction Process

The process of constructing risk devices is not as specific as is hoped. Clear (1988) outlines the formal procedure from which risk devices are constructed, which was discussed in chapter 1. But, this was merely an attempt to explain a complex procedure to general readers who are fascinated by the mysteriousness of the formation of risk devices. In reality, such process is much less formal, prompting designers to make many choices that may be based on research, personal expertise and/or hunches. This section

will seek to unveil the mysteriousness of the construction process while exposing the capriciousness behind many decisions.

The construction of any risk assessment device begins with the availability of large risk data for a specific population from which a wide range of variables and information is collected. Such datasets could contain over 100,000 cases and over a hundred variables. The next step requires the designer to identify the criterion or outcome variable, usually re-arrest, reconviction, or violation of conditions. Because the criterion is usually a dichotomy, pass or fail, a logistic regression with the different variables would be initiated. With hundreds of variables and tons of cases, almost every variable significantly contributes to the variance in the outcome criterion, and the designer must decide on a feasible strategy to reduce the number of predictors included in the risk function. Clearly, a risk assessment instrument with over a hundred predictors would not bode well in the real world where the risk tool would be used. To overcome this obstacle, risk designers would attempt to minimize the size of these risk tools by either selecting risk factors based on prior research or running a regression on all such factors (Baird et al., 2012). In some other instances, the selection of risk factors could be based on whims, some statistical justification, or preference, but, in the end, there is no hard and fast standard to guide the process.

The process of shaving down the number of variables would continue until an acceptable number of factors are removed. The next major decision point involves the determination of the number of risk variables to be included. Again, there is no clear standard that stipulates a specific number of risk variables. It is, however, hoped that the risk tool is practical so that the assessments used with these instruments are not too time

consuming. Others, in an attempt to justify their size preference have argued for parsimony (Flores et al., 2004), while designers of more comprehensive risk tools, such as the LSI, YLS/CMI, COMPAS, or YASI, justify the inclusion of larger numbers of risk items to include both static and dynamic risk factors. Regardless of differences in orientation and preference in the tool's size, designers are confronted with many options.

Next, the selected variables also require a wide range of decisions to determine how each risk item is divided and weighed in the final risk function. A simple perusal of different risk instruments will tell us that risk items could differ by the number of categories into which the information is coded. The designer decides between dividing the information into a dichotomy or into multiple levels. Again, there are different viewpoints regarding this area of decision-making. While dichotomies reduce the amount of time in which assessment staff will spend, increase inter-rater reliability (Austin et al., 2003) and convenience, having variables that partition information into multiple categories also has its benefits. The generic interpretation of standardized coefficients for an independent variable in a logistic regression usually goes something like: for a one-unit increase in "independent variable A", the expected change in log odds is .1563404. What this means is that the relationship rests on a continuum, and additional changes in the "Independent Variable A" would continue to effect changes in the outcome. The benefit, thus, for dividing information into multiple categories is that it will more accurately capture this relationship. Again, the risk device designer must decide on the format that will be embraced.

The score point contribution of each risk item to the total risk function also requires some level of subjective judgment. There are primarily two ways to handle the

problem of weights. First, each risk item has a unique score point contribution to the overall risk function, and the regression computation helps determine the individual weights of each independent variable. Second, some designers have embraced the Burgess Method where risk items are coded either as 0 or 1. The Burgess Method offers convenience to staff, allowing offender's risk scores to be easily computed. The scoring of risk using the actual weights of the risk variables offers increased statistical validity to the risk model, yet it is more time consuming because the summation of scores require more sophisticated levels of addition. The method selected for assigning scores to risk items have enormous implications for the risk model, yet for the purpose now, it is another important decision making point in the construction process.

Finally, risk designers are confronted with the task of determining cutoffs for the different risk categories, which is a two-pronged process. First, the number of risk categories that comprises a risk instrument needs to be determined. Such task should, of course, take into account the different levels of supervision or treatment that are available (Baird, 2009). It would be fundamentally meaningless to divide offenders into five categories of risk, for example, if the agency employing such tool does not offer five levels of supervision that would correspond to the specific needs of each risk group. Despite this reality, many risk designers continue to divide individuals into a range of risk levels irrespective of available supervision or treatment. On the other hand, some experts have also argued for more parsimonious models. For example, Clear makes the argument that risk devices should divide individual into no more than three categories of risk since the different levels of supervisions, even if they were available, may not be meaningfully different (Clear, personal communication, February 28, 2013). Thus, the

number of risk categories to which individuals are assigned, is a task that requires some decision-making.

Second, the cutoffs by which risk groups are delineated needs to be determined. The risk groups should ideally be divided in a manner that maximizes the disparity in their recidivism rate while accounting for group size. The cutoffs points are invisible bounds on the scale score by which groups are separated. Unfortunately, the task of finding suitable and appropriate cutoff points is completely discretionary. Furthermore, this conundrum is often compounded with the lack of a clear and specific standard for determining cutoffs. Once the cutoffs are drawn, the instrument is complete. A final step is to revalidate the risk instrument onto the same population to ensure that the cutoffs are applicable to a different set of cases.

The construction process is based on layers and layers of decisions that may sometimes be informed by good reasons, prior research and sound theories while some are clearly a product of capriciousness. The end result is that many arbitrary decisions eventually amount to a risk device that may have significant consequences for the individual whose risk will be evaluated by the instrument. The difference in cutoff points, though seemingly unimportant in the construction process, may be the fulcrum that determines whether the same individual would have his/her liberty revoked.

Measures of Validity

To date, there exists a wide range of available statistical measures of validity in classification systems (Baird, 2009). Statistical measures of association between

outcomes and risk scores are typically reported using measures of specificity and sensitivity (i.e. receiver operating characteristic (ROC) curve). The ROC curve assesses the accuracy of risk instruments by plotting the true positive rate (sensitivity) and true negative rate (1-specifictiy) for each risk score (Zweig & Campbell, 1993). Thus, the ROC curve represents the range of sensitivities and specificities for a test score (Baird et al., 2012). The Area under the curve (AUC) allows comparisons to be made between ROC curves by using a single measure (Liu et al., 2005).

When three or more risk classifications are defined, the Dispersion Index for Risk (DIFR) is a more suitable measure of risk assessment accuracy then measures that focus on sensitivity and specificity (Silver & Banks, 1988). This is because the DIFR measures potency of a risk assessment by assessing how different risk cohorts are divided by group size and the extent to which group outcomes differ from the base rate for the entire cohort (Baird et al., 2012).

While multiple measures of validity exist for risk assessments, the best overall measure of validity, as identified by Snyder and Gottfredson in *The Mathematics of Classification,* is the level of separation attained in recidivism by risk level when offenders are grouped into risk classifications of meaningful size. Since then, many risk device designers have reiterated the value and practical utility of such measure in measuring validity. Thus, borrowing from the work of Baird and Wagner (2000) and Flores et al. (2006), the current dissertation will measure validity in classification systems by comparing outcome rates for each risk level. Baird (2009) explains that the simple analysis of recidivism rates by risk level should be the standard for evaluating risk classification systems for two specific reasons. First, the plain representation of risk by

subgroup base rates provides clarity to those who use the system; it "conveys more useful information than a correlation coefficient of .25 or an AUC of .70" (p.6). Second, most traditional measures of "predictive validity" place individuals into a yes/no prediction, while classification devices produce a range of risk categories, not just two. It is impractical to use a classification instrument to divide individuals into a dichotomy, especially when decisions in the real world involve a continuum of options (i.e. low, medium, high). Silver and Banks (1998) aptly say, "traditional measures of predictive accuracy, such as sensitivity and specificity, are not the proper way to evaluate the potency of a risk classification model" (p.3). The primary utility of a risk device is in providing a continuum of risk estimates to help guide decision-making. Thus, because the "sensitivity" and "specificity" model only divides offenders into dichotomous categories, its utility is limited in the context of risk classification.

Prior to the discussion about base-rates, it is crucial to make clear that validity in classification is usually not "black or white", but instead stretches across a continuum. Thus, there is not a specific cutoff for what constitutes validity or invalidity; a valid cutoff is therefore context dependent.

Validity in the context of offender risk classification primarily concerns the base-rate estimates of each subclass. To be considered valid, scores should be linearly correlated with the criterion. Ideally, the judicious selection of cutoff scores that define each class (risk level) should produce very different subclass base-rates (Baird, 2009). For example, in an instrument that partitions offenders into three categories of risk (e.g. low, medium, high), the propensity for offending should be starkly different for each

group. Thus, an instrument is deemed invalid when the creation of subclasses does not create large differences in sub-rates.

Table 1 shows validity in the context of cutoff scores and base-rates. The cutoff base rate for each category (i.e. 7%, 17%, and 38%) represents the average propensity for each classification of offender. This risk instrument may be considered more valid because it has achieved meaningful subgroup base rates for each risk category. In other words, the difference among the base rates is large enough to demonstrate that the groups are significantly different for the purposes of programming.

Table 3: RESULTS OF A HYPOTHETICAL VALIDATION OF A RISK SCREENING DEVICE

Scale Score	N	Number of Failures	Number in Cutoff	Cutoff Base Rate (%)
0	**18**	**0**	**123 (37%)**	**7**
1	**16**	**0**		
2	**13**	**1**		
3	**14**	**2**		
4	**16**	**0**		
5	**17**	**2**		
6	**16**	**1**		
7	**13**	**2**		
8	**14**	**3**	**105 (32%)**	**17**
9	**14**	**3**		
10	**16**	**2**		
11	**15**	**1**		
12	**16**	**2**		
13	**14**	**3**		
14	**16**	**4**		
15	**13**	**5**	**105 (32%)**	**38**
16	**14**	**4**		
17	**18**	**5**		
18	**17**	**5**		
19	**16**	**7**		
20	**14**	**8**		
21	**13**	**6**		
Total	**333**	**66**	**333**	

(Adapted from Clear (1988), p. 14)

Next, the propensity for each group (in base-rates) needs to be linearly related to the risk categories in order for it to be considered valid. For example, if the subgroup base-rate for a medium risk category is lesser than the subgroup base-rate for the low risk category, then the instrument is considered invalid because it creates incorrect risk groups. Another indication of invalidity is reversals- i.e., people with a score of 3 failed at twice the rate of those with a score of 11. It is not uncommon for individual risk scores to experience reversals, but it becomes a serious validity issue when the entire risk category

experiences a reversal. For example, if medium risk offenders as a whole failed 7% of the time and low risk offenders failed 17% of the time, the instrument is invalid because the identified low risk group is recidivating at twice the rate of its medium group counterpart.

Validity could also be defined as the ability of a classification instrument to divide groups into manageable sizes for program treatment. The group sizes could be measured by the number of individuals that are assigned to each group. As explained earlier, the fundamental function of classification is to identify low risk individuals and place them under lower cost/lower level of custody for the purpose of relieving fiscal pressures. However, the utility of a classification instrument can be undermined when a large majority of cases are pushed to high risk. When this happens, the agencies may or may not have the proper resources to administer program treatment to this group. Table 2 illustrates the subtle difference in validity between two individual validation studies conducted on Pennsylvania parolees (Austin et al., 2003). To demonstrate that a more concise classification device (Eight Factors from LSI-R) can be more valid, Austin and his colleagues compared the differences in subgroup base rates as well as subgroup sizes. From the study, the "Eight Factors" was declared as a more valid instrument than the LSI-R because of its ability to minimize the number of cases in the high risk category, in addition to having a higher difference in subgroup base rates among the groups. Thus, validity can also be gauged by the sizes of the subgroups that are created, which are fluid depending on the goals of the agency.

Table 4: Outcome comparisons by Risk Level: Pennsylvania Parolees

	Full LSI-R		Eight Factors From LSI-R	
Risk level	N	Rate of	N	Rate of
Low	86 (9%)	Recidivism	146 (15%)	Recidivism
Moderate	398 (40%)	43%	614 (65%)	34%
High	522 (52%)	51%	186 (20%)	53%
		58%		69%

(Adapted from Austin, Coleman, Peyton, & Johnson (2003).

Validity in classification devices can be measured by comparing subgroup base rates and number of cases assigned to each risk category, and whether classification outcomes suit the intended goals. The difficulty in gauging validity comes from the lack of a clear definition for what constitutes validity (Baird & Wagner, 2000). Validity is measured on a continuum, where a precise standard is often non-existent. Furthermore, this range in validity should be reflected in a healthy balance between base rates and number of cases with the intended goals, to optimize each without undermining the other. As such, the determination of validity and the construction of classification devices are an art, rather than a perfect science.

Evolution and Validity

Historically, the emergence and evolution of risk assessment transpired alongside the needs of our criminal justice system. This is because the development of the criminal justice system is heavily intertwined with risk assessment to ensure that the system as a whole functions efficiently (Brenan, 1987). The origin of risk assessment can be dated back to 1870 to Cesare Lombroso's time when he tried to identify and categorize people's propensity for crime based on physical attributes that resembled the primitive man (Lombroso, 1876). Similarly, in the 1900s, Goring studied the physical and psychological attributes of people with known predisposition to crime (Gottfredson, 1987). Though these methods were both unscientific and atheoretical, they represent some of the earliest attempts to classify people.

1st Generation

Although risk assessment technology was first introduced 80 years ago (Burgess, 1928), it expanded most rapidly over the recent decades. Bonta and Andrews (2007) described these advances in terms of four generations of risk assessments. The first generation of risk assessments was known as clinical judgments; they were based on assessor's intuitive judgment or gut feeling. On the basis of clinical judgment formed through their knowledge of the case, their understanding of criminal behavior and their experiences with similar offenders, correctional specialists would attempt to predict problem behaviors (e.g., Historical, Clinical, Risk Management, HCR-20; Andrews, Bonta, & Wormith, 2006). This generation of prediction represents the earliest attempts to classify offenders for criminal justice purposes- the first recorded endeavors towards

offender classification and placement can be traced back to France in the mid 1700's. Contrary to the overcrowded prisons elsewhere, in *Maison de Force* (House of Enforcement), inmates were lodged in separate quarters, adequately clothed, and well fed- all made possible because they had implemented their own classification system (Champion, 1996). However crude these early classification schemes were, at the time, it represented pioneering events and had auspicious consequences for inmates.

Today, clinical predictions are deemed outdated, primitive, and unscientific because it relies on the objectivity and skill of the specialist making the risk predictions. Concomitantly, meta-analyses consistently find them to be inferior to newer mechanical methods in the prediction of clinical outcome and dangerousness (Bonta, Law, & Hanson, 1998; Mossman, 1994). Averaged across six 1st generation mean estimates, the overall mean r was .12, meaning that predicting failure or recidivism was only accurate 12 percent of the time (Andrews, Bonta, & Wormith, 2006). Because of its inaccuracy, today most correction facilities use more modern and mechanical prediction techniques that have proven reliability and superiority over clinical judgment. However, this does not mean that clinical judgment is completely obsolete. In some instances today, psychiatrists and psychologists with extensive clinical training and experience with deviant conduct and criminal behavior still prefer clinical predictions (Champion, 1994). Yet despite their best attempts to clinically predict, accuracy remains an issue. Moreover, clinical predictions are costly compared to more advanced methods since each clinical prediction is individualized (Clear, 1988). However inferior to future methods of prediction, fundamental clinical judgments spurred decades of research that allowed for major advancement in risk prediction.

2nd Generation

Beginning in the 1970's, there was a specific aim to increase accuracy of risk prediction- a new era was borne. It revolutionized risk prediction by attacking the problem elements of clinical prediction. Humans are subjective, unreliable, and highly susceptible to blundering. Thus, this new era of risk prediction eradicated the human calculation of risk. Second generation risk assessments (e.g., Salient Factor Score or SFS; Farrington & Tarling, 1985) were actuarial based and considered socio-demographic factors (e.g., age at first arrest, employment, and drug history) and criminal variables (e.g., number of convictions, parole history, and types of offenses) that have been demonstrated to increase the risk of reoffending. It assigns these items quantitative scores which can be summed- the higher the score, the higher the risk that the offender will reoffend. Meta-analytic studies show 2nd generation assessment instruments outperform 1st generation clinical prediction; overall mean r were.42 and .12 respectively (Andrews, Bonta & Wormith, 2006).

The biggest aspect of change in actuarial instruments is that it specifically limits the discretion of decision-makers by stipulating that evaluations be made based on the risk variables that are proven to be significantly correlated to risk or recidivism. However, despite this attempt to remove human discretion from this equation, some discretion is still necessitated by the instrument. When the information obtained from the interviews with inmate and past history records are transcribed onto the actuarial risk assessment instrument, human discretion is inevitably needed.

Second generation risk assessment relies solely on static (i.e., historical, unchangeable) factors such as criminal history, which some criminologists argue does not align with the rehabilitative goals of corrections. Criminal history and other factors that sample past behavior are treated as static risk factors. According to Bonta and Andrews (2007), this poses a major shortcoming for second generation risk assessment because the scales do not account for offenders changing. Future advancement in risk prediction technology incorporates needs factors to account for change in offender's predisposition towards reoffending (to be discussed). However, there is much debate as to whether its refined successors (3rd generation and 4th generation) can achieve higher accuracy in prediction when using more technologically advanced instruments. Interestingly, meta-analysis shows that 2nd generation risk assessment instruments predict better than 3rd and 4th generation risk assessment instruments; overall mean r are .42, .38, and .41 respectively (Andrews, Bonta, and Wormith, 2006). This begs the question of why corrections are willing to adopt more advanced risk assessment instruments when their reliability is known to be less.

3rd Generation

Static risk assessment can be useful for classification purposes but are constrained by an inability to contribute to the effective treatment planning and ongoing evaluation of offenders. Third generation risk assessments (e.g., Psychopathy Checklist-Revised or PCL-R, Violent Risk Appraisal Guide or VRAG, Self-Appraisal Questionnaire or SAQ; Loza & Loza-Fanous, 2001) are risk/needs assessments that include dynamic risk factors (e.g., criminal attitudes and companions) and static factors from 2G instruments. Dynamic questions were asked about present employment, criminal friends, family

relationships, etc. to help decision makers gain insights on the offenders' current and ever changing situation. Evidence suggests that changes in the scores on some of these risk-need instruments correlate with changes in recidivism (Andrews & Robinson, 1984; Motiuk, Bonta & Andrews, 1990).

Third generation risk instruments are sensitive to changes in an offender's immediate circumstances. It also provides correctional staff with information as to what needs should be targeted in their interventions/incarceration. One decisive advantage 3G assessments have over its predecessors is that they are particularly useful in guiding the delivery of rehabilitation services and measuring change, which is often a major focus of correctional agencies (Bonta & Andrews, 2007).

However, multiple meta-analytic studies have shown that 3rd generation risk instruments are weaker in predicting recidivism than 2nd generation risk instruments (Gredreau, Goggin & Smith, 2002; Hemphill & Hare, 2004; Andrews, Bonta, & Wormith, 2006). The most widely used 3G instrument, the Level of Service-Revised (LSI-R), achieves an overall r of .36, which is less than the mean r for 2G risk instruments (Gredreau, Goggin, & Smith, 2002) . The best predictors of recidivism are static variables, which are comprised of criminal history and sociodemographic variables. Adding dynamic variables to actuarial instruments is a valiant effort to assess and treat the needs of offenders, but from a prediction perspective, it provides noise that weakens the instrument's ability to predict risk (Austin et al, 2003).

Despite this knowledge that 3G instruments are less accurate that its predecessors, virtually all correctional facilities have supplanted the older static tools with 3G

instruments. This conflict of interests illuminates on the public's willingness to have risk instruments align with rehabilitative ideals, even when it goes against a fundamental goal of risk prediction, which is to accurately classify inmates. This is a clear sign that the goal of corrections has shifted from accurately classifying inmates to attending to the needs of inmates.

4th Generation

Finally, the last few years have seen the introduction of fourth generation risk assessments, which place more emphasis on rehabilitation. This new assessment instrument integrates systematic intervention and monitoring with the assessment of a broader range of offender risk factors (e.g., Level of Service/Case Management Inventory or LS/CM). Andrews et al. (2006) showed that the LS/CM outperforms 3G instruments, overall r were .41 and .36 respectively. However, the validity of 4G risk assessment instruments is understudied; literature and evaluations of LS/CM are still lacking and the use of it is in its fledgling stage.

The one study by Andrews, Bonta, and Wormith (2006) should not be taken as anything conclusive. Research methodologists argue that results obtained in this study could have been confounded in ways that did not exist for earlier instruments (2006). For example, because there is a constant interaction between treatment outcome and risk/needs assessment, the training, experience, and clinical supervision of users are important moderators of predictive criterion validity. Fourth generation risk instruments requires more human discretion than 3G instruments to function properly, which, as discussed earlier, increases risk of subjectivity and bias. Furthermore, in research

methods, confidence in treatment effect is a function of knowing the exact pre-treatment and post-treatment scores. The simultaneous treatment and evaluation of risk/needs may greatly reduce the risk/needs scores if offenders were appropriately treated, which might potentially confound treatment effects. Thus, until more evaluative studies are conducted, further analysis is needed to know how the new 4G instruments compare against older instruments.

Rationale/Validity for Each Generation

The popularity for each of the four generations of risk assessment was caused by a confluence of factors. While the notion of redesigning risk devices to increase validity and accuracy has always been the propelling force from which change is wielded, there have been times when agencies lose sight of this ideal, allowing evolutionary takeovers and shifts to be precipitated by ancillary goals. This section will seek to understand the rationale for the shifts by objectively analyzing the multitude of factors that play a strong role in molding such ideological shifts.

1Gs and 2Gs

First generation risk assessment or clinical judgment was born solely out of the necessity to sort masses of inmates with varying degrees of risk in a humane manner. In England prior to the introduction of clinical risk assessment, prisons often celled large numbers of inmates, including males, females, and children in deplorable conditions (Champion, 1994). Similar conditions were found in prisons in the United States, France, Scottland, and countries throughout most parts of Europe (1994). The discovery of risk assessment paved the road for major prison reforms, which transformed large

unmanageable masses into groups of inmates with varying needs and proclivity towards violence.

Today, meta-analytic studies tell us that 1G instruments were crude attempts to estimate risk; their overall ability to accurately assess risk is low, average mean r is .12 (Bonta, Low, & Hanson, 1998; Hanson & Morton-bourgon, 1998). Despite the fact that estimating risk with clinical risk assessment methods were only slightly better than chance, it was the only alternative available at the time.

Two primary philosophical orientations, science and utility, helped push for changes in risk prediction technology in corrections. Science ensures objectivity. Instead of treating offenders as individuals, a scientific model treats offenders as members of groups based on experiences of other members of the group to which they belong (Clear, 1988). It asks questions of how an offender is similar to others that have been experienced in the past. Thus, science allows decision-makers to objectively treat people based on classes of past experiences; predicting future incidences of crime is no longer an intuitive judgment based on a decision-makers whimsical prognosis. The increased certainty of 2G predictions come from the ability to systematically and scientifically objectify risk based on decisions on what is known about human behavior.

Another important aspect of science is certainty. By employing methods that are derived from statistical analysis of the past predictors of risk or recidivism, 2G instruments significantly improved over 1G clinical methods (Farrington & Tarling, 1985). Actuarial/mechanical strategies improved the predictive validity of 2G instruments over its predecessors by more than fourfold, from mean r = .10 to r =.40.

Clinical predictions from 1G instruments were proven to be inferior because it did not take into account the predictors of risk on an aggregate level- the experiences of a decision-maker is limited to only his/her own experiences. Thus, increasing certainty is a major scientific goal of risk instruments.

Of course, to say that any risk assessment instrument can be completely confident in predicting future human behavior is a lie. It is not that the values of science are wrong, but the quality of the science of human behavior is very limited (Clear, 1988). Science has been extremely instrumental in improving risk assessment validity. It however cannot predict human behavior with 100 percent certainty. And this is the heart of the problem that needs to be considered when important decisions about a person's liberty are at stake.

The second philosophical assumption is that it is appropriate to make decisions based on what offenders will do in the future. The use of prediction methods in correctional decision-making is considered utilitarian because its aim is the design of a punishment level or type that is best able to reduce the incidence of future crime. Prescribing a severe punishment for someone who will never commit another crime is excessive and prescribing a "slap of the wrist" type punishment to a high-risk individual undermines justice, especially when the individual quickly commits another crime upon release. Through the widespread use of more accurate prediction tools, there will be a greater reduction in harm, pain, and suffering in a system that is predicated towards treatment, incapacitation, and specific deterrence. Thus, the underlying purpose for making punishments more utilitarian is justice, which is the foundation of the United States criminal justice system.

Science and utilitarian punishments are the philosophical basis that fueled the popularity of actuarial risk assessment. However, other events transpired at around the same that helped actuarial risk assessment instruments grow in pandemic proportions. During the late 1970's, 1980's, and into the 1990's, the prison population mushroomed in the United States, mainly due to stricter laws (Voorhis & Brown, 1996). According to Blumstein and Wallman (2006), the incarceration rate by 1999 was over 4.3 times the rate that had prevailed for the earlier fifty years. Coupled with the ebb in correctional spending, the criminal justice system faced very serious issues (2004). Criminologists searched for practical solutions and found relief from risk prediction- the use of prisoner classification helped free up resources by allocating intensive treatment only to the most serious offenders (Clear, 1988). Inmates who posed less threat to others and society either made early release from prison or were lodged in minimum security facilities, which cost significantly less than their more secure counterparts. Maximum security prisons could cost up to $70,000 dollars for each inmate, while minimum security facilities cost only a fraction of that (Rhodes, 2004). Thus, risk prediction provided an equitable and fair way to alleviate fiscal pressures during a time when alternative solutions were lacking. The popularity of 2G risk assessment instruments was caused by the prison climate at the time along with the philosophical basis of science and utilitarian punishments.

3G and 4G

Over the last two decades, offender rehabilitation made a powerful comeback. Large scale meta-analytic studies from Canadian Corrections, Tong and Farrington (2006), and Sherman et al (1997), consistently found a strong link between rehabilitation

and effective reduction in reoffending. Following this trend in corrections, logically, criminologists began to align policies with rehabilitation. Risk prediction was not an exception. Criminologists quickly attacked 2G risk instruments for failing to account for the personal needs of each inmate. The static factors from 2G instruments focused on demographics, prior arrest records, and other background characteristics that did not change, meaning an offender's predisposition towards offending would never decrease (Bonta & Andrews, 2007). On the other hand, 3G risk instruments or risk/needs assessments included dynamic variables such as substance abuse, employment, and criminal companions. Changes in risk scores signal changes in the likelihood of committing a new offense, unlike static factors. This is important for correctional programs and the staff charged with managing offender risk.

However, there is much evidence suggesting that the creators of 3G risk instruments have forsaken the fundamental principle of risk assessment, which was to validly divide individuals into risk categories. The goal to couple risk assessment with needs evaluation has added to the length and complexity of many risk assessment instruments (Baird, et al., 2012). The inclusion of dynamic risk factors weakens the ability to predict risk because such risk factors are statistically irrelevant to most outcome criterion (Baird, 2009). Recent meta-analyses supported the allegation that 3G risk tools make weak predictors of risk. In evaluating 47 studies of LSI validity, Vose, Cullen, and Smith, (2008) found that there was substantial variance in the correlations obtained, and coefficients as low as .137 were cited. In another meta-analysis that evaluated 22 LSI validation studies, Campbell, French, and Gendreau (2007) noted that the average correlation between LSI scores and recidivism was .24. Hence, instead of trying to

increase certainty and accuracy in prediction, 3G instruments detract from such goals. Furthermore, the addition of dynamic variables injects noise in the instrument because it requires more human involvement, which opens up more opportunities for bias and subjectivity to take center role. By making risk assessment instruments more "needs" focused, it came with heavy heavy tradeoffs and consequences, it dismissed decades of research and advancement in accurate risk classification. This dilemma begs the question of whether following a "needs" trend in rehabilitation is more important than justice and fairness, because, essentially, less accurate risk assessment instruments will produce more erroneous classifications of inmates.

In retrospect, 3Gs represented a shift from the scientific risk prediction model to one that fused rehabilitative ideals with risk reduction. The rationale for 4G risk instruments is to bridge this disjuncture between risk assessment and intervention. The idea to align risk/needs instruments with intervention is not new, and had existed long since 3G. Until 4G tools were conceptualized, however, they remained separate entities without a shared unified goal. The strength and rationale of 4G instruments come from the fact that both historically detached systems were now merged to serve one purpose. Furthermore, fourth-generation tools added the concept of responsivity, which was intended to measure both an individual's readiness for change and the offender's ability to respond to particular treatment programs (Baird et al., 2012). The evolution of risk assessment demonstrates how rehabilitation has regained its footing in contemporary American corrections.

Punishment based on needs is a problem associated with the 4th generation risk paradigm. As explained by Clear (1988), punishment based on risk is justified by which

actual crime reduction is experienced. Ideally, a potential offender will be less harmful to society if he is confined- the benefits here are obvious. However, as risk instruments incorporate more needs variables, the reason for punishment in corrections becomes synonymous with punishment based on needs. For example, early release in parole may be revoked if an offender has no stable housing, job, or supportive family members awaiting his release. Traditionally, this decision to revoke parole would be based on offender's proclivity towards reoffending, but now, it may be revoked based on lack of support in the community.

Third and fourth generation risk tools were based on several developing ideologies, and were no longer predicated on increasing validity. In other words, unlike their earlier counterparts (e.g. 1G and 2G), whose technological advancements were born primarily out of creating valid, effective, and accurate classification systems, 3G and 4G risk tools aspired towards other ideologies. Third generation instruments were based on measuring needs, while 4G instruments were based on aligning case management with risk designation. As such, 3G and 4G risk instruments, such as YLS/CMI YASI, Positive Achievement Change Tool (PACT), SAVRY, and COMPAS, did not employ standard actuarial methods of development. There is no construction sample from which the tools were built, but instead, these risk tools incorporate risk factors identified in prior research studies and are based on one or more theories of criminal or deviant behavior (Baird et al. 2012). The standard method of risk device construction (described in chapter 1) ensures that the risk factors selected are optimally related to risk. Without this procedure, the risk factors would not be selected based on their statistical relationship to the outcome. Baird

and colleagues (2012) argues that 3G and 4G instruments have lower validity because they don't employ this standard method of risk device construction.

Furthermore, it appears that the selection of risk factors was based very loosely on prior research, which would also explain the reduced validity in 3G and 4G risk instruments. Baird and colleagues (2012) attribute this reliance on theories to the result of tying development to a particular type of crime. They found that "many of the factors added to generation 3 or 4 models have little statistical relationship to recidivism" (Baird et al., 2012, p.4). This fact is not fully appreciated because relatively few published studies of these risk tools included individual item analysis. Recently, however, some experts suggested that the removal of statistically irrelevant variables would drastically increase the validity of these risk tools (Flores et al., 2004; Austin et al., 2003). The general pursuit to incorporate 3G and 4G ideologies decreases an assessment system's ability to accurately identify high-risk offenders while simultaneously compromising all other assessment objectives.

So why are these risk instruments ubiquitously adopted by so many agencies? Baird and colleagues (2012) explain that the widespread adoption of 3G and 4G instruments has to do with: 1) overselling of the evidence behind these risk models, 2) improper comparison of validation studies to undermine 2nd generation risk tools, 3) misrepresenting various revised models of the Wisconsin system as a "static" risk tool to justify the need for change, and 4) employing questionable methods of comparison. Thus, the combined effects of misrepresentation and research blunders launched the widespread use of 3rd and 4th generation risk tools.

Reliability

All studies in risk assessment emphasize the need for reliability/consistency among decision makers. Without consistency, decision-making is inevitably weakened. Such disparities are a good indicator that human subjectivity and bias are subverting the goals of structured risk assessments. Thus, it is important for risk assessment tools to demonstrate high levels of inter-rater reliability. In other words, given the same information and facts about an individual, decision makers should make the same decision consistently, irrespective of their differences, subjectivity, and preferences.

Studies on risk assessment instruments found that different risk properties and different social elements are critically important for ensuring inter-reliability. Reliability is particularly critical when models include 25 or more risk items, which often require subjective judgment (Baird et al., 2012). Static variables, compared with dynamic ones, are more consistently rated because they require less subjective judgment (Austin et al., 2003; Baird, Heinz, & Bemus, 1979). This is because static variable are often found from an individual's official record. Dynamic risk variables seek to measure family relations, residential stability, and criminal friends, which require greater levels of subjective judgment. Risk tools that embrace the risk/needs model, such as the PACT, YASI, and LSI include many more dynamic variables, making them more susceptible to reliability problems. Thus, the surest way to ensure reliability is to increase consistency among staff completing risk assessments.

A multitude of social factors have also been found to impact reliability. If staff members administering the risk assessment instrument take on a particular liking towards

a risk tool, there would be higher levels of inter-rater reliability. Haas and DeTardo-Bora (2009) found that staff's negative attitude towards a risk tool will increase tendencies to override the tool's classification decisions with their own decision. Also, Lowenkamp and colleagues (2004) found that higher levels of staff training are critical in augmenting reliability. Hence, many social factors also affect an instrument's reliability.

Given the vast knowledge and understanding that exist, the issue of employing risk assessment tools with low reliability is completely preventable. Unfortunately, in some instances, risk assessment models are marketed before any reliability analyses are conducted (Baird et al., 2012). In other instances, inappropriate measures of reliability are used. For example, simple correlations were used to estimate inter-rater reliability (Andrews, 1982; Andrews & Robinson, 1984; Rettinger, 1998) despite strong evidence suggesting that correlations only measure patterns between raters and not necessarily agreement. Baird and colleagues (2012) posit that it is theoretically possible to have high levels of correlations even when two raters never agree.

To remedy this problem, intra-class correlation (ICC), which accounts for the difference in magnitude for different ratings, can be used. However, the problem that plagues this measure is that it fails to be an accurate estimator of reliability when different actions and decisions result. As such, Baird (2009) argues that reliability analyses should only be conducted using percent agreement because it is more accurate and it more clearly conveys information that is important to case decision-making.

In more extreme cases where inappropriate measures of reliability are used, the focus on the level of "internal consistency" has been cited. Such measure is adopted

from the psychology field, and makes a better measure of constructs, rather then reliability in risk assessment instruments. Cronbach's alpha, which is the measure used to evaluate internal consistency, is a competent estimator of how well responses correlate with each other (Garson, 2003). This measure is valuable when measuring psychological constructs like depression, happiness, or emptiness, for which there is no litmus test. As for evaluating the reliability of risk tools, recidivism is not a construct, and can be measured directly.

Equity

It is critical for any risk assessment processes to treat all subpopulations equitably, as well as reduce the inequities that plaque the criminal justice system to the extent possible. In a narrower sense, equity in offender risk assessments refers to the judicious selection of variables that would otherwise pointedly discriminate against specific groups. Thus, specific variables that pertain to race/ethnicity, gender, jurisdiction, and socioeconomic status are typically excluded from risk tools. The pursuit for more equitable risk tools, often times, involves the removal of valid and robust predictors unfortunately.

The goal to ensure equity in risk assessment processes presents a dilemma. Especially, risk tool designers are torn between including inequitable risk variables and omitting strong yet, discriminatory predictors. On the one hand, such risk variables would vastly augment the capacity to classify individuals, but on the other hand, issues regarding ethics are raised when risk tools begin to focus on specific demographic characteristics that border between right and wrong. The inclusion of such variables can

significantly enhance an instrument's ability to classify risky individuals. For example, Van Voorhis, Salisbury, Wright, & Bauman (2008) suggest that separate instruments may be required to optimize classification results for girls, mainly because gender has such a strong relationship with risk. Similarly, offenders from different racial/ethnic backgrounds often have very different recidivism rates (Tonry, 1987). In other words, race variables make a strong and robust predictor of reoffending outcomes.

The locations in which offenders are caught and arrested are also strongly linked to risk and recidivism. In fact, it is common knowledge among risk tool designers to control for locations when constructing risk assessment instruments. Any risk construction process that does not control for jurisdictional disparities would inevitably create risk devices that are biased against specific locales. Furthermore, such differential impact on an individual's risk level is suggestive that social factors, such as police efficiency, play a pivotal role in determining risk. The inclusion of geographical variables would be inherently wrong because such forces are beyond the control of individuals. Thus, the risk of recidivating, with all things being equal, can change starkly from one location to another, supporting the perspective that geographical variables should not be used in determining risk.

The need to treat all subpopulations equitably seems obvious. However, there is also overwhelming evidence suggesting that the inclusion of such demographic characteristics would greatly enhance the validity of risk tools. Berk and colleagues (2009), who used "random forest modeling" to build hundreds of different risk models that utilized all available offender information, were able to augment validity of risk tools by suspending the restrictions that were typically levied against risk designers during risk

tool construction processes. In including every potential predictor variable in their risk analysis and tool construction process such as geographical locations, specific personal characteristics, and even "shoe size", one of the most robust classification instruments the risk world has ever witnessed was built. Of course, much contention has once emerged due to the nature of these variables, the applicability of such tools, and whether it is ethical to target specific subpopulations.

Risk designers are ultimately torn apart by such competing standards, presenting a major paradoxical complication. Specific demographic variables are extremely predictive of future risk, but the inclusion of such information in risk tools often borders between right and wrong. However, Baird et al. (2012) cautions that the problem of equity is not typically addressed in validations or evaluations, and that risk instruments should not be implemented before equity is firmly established.

Cost and Efficiency

It is completely necessary for risk tools to be cost effective and efficient. There are two types of cost/benefit analyses associated with risk tools. First, there are costs related to development time, staff training, and technical assistance, all of which could substantially drive up the cost of implementation (Baird et al., 2012). Such analysis should also consider training costs, staff time, and travel expenses incurred in attending training. And finally, the cost analysis should include an estimate of agency staff time required to fully implement the model, and the information technology costs incurred. The best risk models would not be deemed useful if their implementation costs are exorbitantly high and unaffordable.

Risk tools could vary greatly by their implementation costs. However, it is difficult to objectively compare such costs across different risk tools because the size of the agency and facility employing such tool plays a big role in determining cost. For example, a recent evaluation of implementation costs by the Baird and colleagues (2012) found that the implementation of the PACT in Florida and Georgia cost roughly 1.2 million dollars over seven years. And the YASI implemented in Virginia cost considerably less, roughly 100,000 dollars. The criminal justice field will have to determine if the additional time required to complete different risk models produces enough added benefit to justify the cost incurred.

Second, the length of time required for risk staff to complete and perform a risk assessment on an individual is another critical factor related to cost and efficiency. Some risk assessment procedures could require two or more hours to complete. Spending such lengthy time on risk assessment would be justified if the assessment leads to better classification, better decisions regarding placement and services and better outcomes (Baird et al., 2012). One could make the argument that if staff were given unlimited resources and time to conduct a full-blown investigation about an individual's past and risk factors, the validity would be greatly enhanced. Clearly, that wouldn't be feasible in a correctional setting where thousands of assessments are routinely made. However, if a risk model's capacity to classify is not substantially enhanced, then the time required to complete these systems would not be warranted, especially if the agency is already under-resourced.

This chapter was an attempt to explain the competing standards used to guide the construction and evaluation processes of risk tools. The difficulty in building a working

risk model is therefore complicated as risk designers are pulled into different directions, trying to satisfy different requirements in validity, reliability, equity, and cost. The decision making processes and reasoning used to reconcile these opposing ideologies have much impact in shaping a risk assessment instrument.

Chapter 4- Methodology

Statement of the Problem

Few attempts to evaluate the sensitivity of error in risk devices have been made. In particular, the problem of how errors from: official records, staff, and the misuse of instruments impact the effectiveness of offender risk classification is understudied. Furthermore, it is not known how different risk device properties can increase or decrease the sensitivity of error in risk instruments. Thus, the study models the impact of errors in risk data and information on the overall validity of classification instruments. The sensitivity of error for every risk device is different depending on specific risk instrument properties such as case distribution, cutoffs, number of risk categories. Such properties directly impact the tolerance and sensitivity of error in risk instruments. This dissertation argues that the number of final misplacement of individuals will depend on two essential elements: quantity of error and sensitivity.

This research inquiry is answered by using both conceptual and actual data. Risk data and instruments that would facilitate the testing of all the proposed research questions and hypotheses are not readily available. This is because, in order to explore the different facets of the proposed inquiry, specific situations are requisite- and these particular situations may be easier tailored into a fabricated data than to be found in the real world. The sample is engineered using Monte Carlo Studies: simulation methods making use of random draws from an error distribution and multiple replications over a set of known parameters. This methodology is particularly relevant in situations where the only analytical findings involve asymptotic, large-sample results (Mooney, 1997).

Many sophisticated statistical software today such as STATA, SAS, and SPSS come standard with the "random number generator" function. STATA is used for the construction of data, and is generally preferred over other similar programs for completing the desired tasks. For each possible statistical scenario in question, the random number generator function will be used to draw the necessary random numbers to create sets of risk data that would be similar to those found in the real world of offender risk assessments.

Datasets

To answer every proposed research question, several risk assessment instruments are used. Two main datasets, *Risk Device X* and *Oregon JCP FIRE,* provide the necessary tools to test many of the hypotheses. While *Risk Device X* is based on conceptual data, *Oregon JCP FIRE* is an actual risk tool currently used to classify risk in Juveniles. Both of these datasets are manipulated to form additional subsets that would facilitate answering all of the other research questions. Deriving from *Risk Device X*, *Risk Device 5 Cat* is created. Also, various versions of the *Oregon JCP FIRE* are built, which includes: *Oregon JCP 3 Cat, Oregon JCP-Burgess, and Oregon JCP-Coefficients.* Their method of construction, as well as their strengths and weaknesses are discussed in their respective sections.

Risk Device X- Dataset Construction

One of the two main risk datasets/instruments, Risk Device X, is engineered using Monte Carlo Studies: simulation methods making use of random draws from an error distribution and multiple replications over a set of known parameters. The cutoff scores

are determined before implementing Monte Carlo simulations. In reality, the cutoff scores for each risk category depends largely on 1) the instrument, 2) the number of items it has, 3) number of risk classification categories there are, and 4) the discretion of local jurisdictions and agencies, who are encouraged to develop their own cut-off scores which conform to local norms and needs (Andrews & Bonta, 2003). Once the numeric boundaries for the cutoff scores are determined, the random number generator function can be used. The cutoff scores are an important prerequisite because it will provide the lower and upper limits for each random number draw. For example, the cut-off scores developed for the original offender population on which the LSI-R was validated were 0-13 for low risk/need, 14-23 for low-moderate risk/need, 24-33 for moderate risk/need, 34-40 for medium-high risk/need, and 41-47 for high risk/need. Thus, the boundaries for the random number draws will follow this set of cut-off scores. For simplicity, the numbers from which the random draws will be taken will range from 0 and 100 and the subgroup cut-off scores will be created by dividing 100 by the number of risk categories. Thus, if there are 5 risk categories, respectively, the cutoff scores will be 1-20, 21-40, 41-60, 61-80, and 81-100.

For the purpose of manageability, 1,000 hypothetical cases were generated and forced into various distributions in question- into normal, positively skewed, negatively skewed, platykurtic and leptokurtic frequency distributions. The specific characteristics for each individual case, here, are not very meaningful, but analyzed collectively, the cases will offer a chance to visualize and understand the interplay of error in the risk data and error in classification outcomes. Thus, the construction of each case on the individual level is ancillary to the construction of collective distributions within these

samples. The primary objective is to understand how the combined differences and similarities within the 1000 cases will influence sensitivity. Thus, minute differences among the individual cases, particularly variables like gender and race, are ignored. As well, the type of risk items or the kind of risk information that would normally set the distinction between risk variables is ignored. As such, *Risk Device X* contains only 11 risk items that are essential to the overall risk function. Table 1 shows the distributions of the cases on an aggregate level for all 11 risk items including: dichotomous risk items, 3 leveled risk items, and 4 leveled risk items, which are specifically designed to facilitate the testing of many of the study's hypotheses.

Next, the specific procedure for constructing *Risk Device X* is explained. Prior to implementing this procedure, several key elements were considered. The risk device contains 11 risk items with different distributions and categories (see table 1), and there are 1000 individual cases altogether. The steps are outlined below.

Step 1: Create 1,000 cases with 11 variables or items. At this point the seed should be set so that the same random number draws can be replicated in the future.

Step 2: Next, random numbers between 0 and 1 are generated for each variable and case.

Step 3: Before executing this step, it is important to understand the general distribution of cases across all the variables so that the upper and lower limits for each risk category can be determined. The numbers for each variable are then recoded from a continuous variable (0 to 1.0) to an ordinal one (1, 2, 3) based on

the cutoffs. In terms of the actual risk instrument, this step seeks to replicate the creation of different options for a risk item.

Step 4: A critical attribute of this simulation is to replicate different statistical scenarios that were posed in "research questions." This step will re-execute step 3 so that different research scenarios are properly imposed onto each variable. For example, if "item H" is being replicated, the 1,000 random cases need to be divided so that it follows the desired schematic (100, 200, and 700). In other words, the upper and lower limits for the three options need to be set based on the desired distribution. For the creation of each risk item, these steps will be taken.

Step 5: Once the 11 risk items or variables are created using steps 1-3, the risk classification designation for each individual can be determined. If a risk instrument has three risk levels or categories, the scores for the individuals can be combined to form the following schematic: (0 1 2 3 = 1) (4 5 = 2) (6 7 8 = 3). From this step, all the specific properties are built into each scenario to answer the proposed research questions.

Step 6: The final step relates each individual case to their respective base rate. To accomplish this, first, a new variable needs to be created. Because the targeted base rate for the group is 40 percent, 400 of the 1000 cases are set to fail. However, unlike many of the earlier steps where the random number generator is used, the cases to which a failure status is assigned cannot be randomly generated, because a linear relationship between the risk scores and recidivism is expected. Specific steps are taken to insure that these parameters are met. First, the cases

need to be placed in a numeric order based on their total risk score, with the cases in the earlier segment having a lower risk score than latter cases. The base rate is then assigned to each case based on its alignment with risk scores. As we go down the list of cases, risk scores generally exhibit increasing patterns. The base rate or the number of failures should also progressively increase while going down the list. After 400 cases of failure have been completely assigned to the 1000 cases, a logistic regression is run on all the risk items against the criterion variable. When the pseudo r square is .40, we would have met our mark. In addition, the researcher must check to confirm that too much multi-collinearity doesn't exist between any two risk items, or exceed an r of .2.

The construction phase is complete when all the following distributions with different risk categories are imposed onto the dataset.

The full description of how the individuals in Risk Device X are divided into risk categories and into their respective subgroup base rates can be seen in the following table.

Table 5: *Description of Risk Device X*

Items	Example	Categories	Distributions			
A	Criminal History	0, 1, 2, 3	50	150	250	550
B	First Arrest	0, 1, 2, 3	550	250	150	50
C	Attitude	2, 4, 6, 8	250	250	250	250
D	Residential stability	0, 1, 2, 4	150	350	350	150
E	Criminal Friends	0, 1	500		500	
F	Last Job- Duration	0, 1	800		200	
G	Education	0, 1	200		800	
H	Marriage	0, 1, 2	100	200	700	
I	Substance Use	0, 1, 2	333	333	333	
J	Family Poverty	2,4,6	600	300	100	
K	Need	0,1,2	800	100	100	

Table 6*: Scale Score for Risk Device X*

Scale Score	N	Number of Failures	Number in Cutoff	Failure Rate (%)
7	2	0		
8	2	0		
9	3	0		
10	14	1		
11	31	2	276 (27.6%)	(7.2%)
12	57	2		
13	74	5		
14	93	10		
15	102	12		
16	119	12		
17	109	22	330 (33%)	(13.9%)
18	105	75		
19	90	80		
20	74	64		
21	43	37		
22	39	36		
23	23	21	394 (39.4%)	(84%)
24	13	11		
25	6	6		
26	0	0		
27	1	1		
Total	1000	397	1000	

Critique

To facilitate the testing of various hypotheses, *Risk Device X* is specifically engineered to be eclectic and free-ranging. It contains risk items with very different skews, i.e. platykurtic, positive, negative, and normal. Also, the instrument is comprised of risk items that are dichotomous, three-leveled, and four-leveled. However, in reality, risk assessment instruments are prosaic and made up of uniform risk items that are only narrowly dissimilar. For example, the Oregon JCP FIRE contains 30 risk dichotomous risk items, most of which are positively skewed. Likewise, the LSI-R is comprised of 54

risk items that are mostly dichotomous. Thus, one downside of creating a risk tool that could potentially answer all of the research questions is that is lacks generalizeability.

Risk Device X contains cases that are strongly correlated to risk, which is generally not found in the real world of risk assessments. Simply put, Risk Device X is an outstandingly valid risk classification instrument. The separation in subgroup base-rates achieved among the three risk groups is exceedingly high. For instance, the low risk group recidivates at a rate of 7.2%, the moderate risk group recidivates at a rate of 13.9% and the high risk group recidivates at a rate of 84%, rendering it one of the best classification instruments in existence. Thus, the interpretation of any analytical findings using *Risk Device X* requires some level of discretion.

The strength of using *Risk Device X* is that it shows an enormous capacity to separate individuals into groups with substantially different subgroup recidivism/base rates. In fact, it was the specific goal of the construction process to build a close-to-perfect risk model to which, then, error could be injected. The need to construct a conceptual model that demonstrates such capacity to classify individuals is born out of the fact that any actual risk instrument/dataset is already tainted with error in the risk information. *Risk Device X,* in this sense*,* is the complete opposite because it is arguably uncorrupted by error; it epitomizes the best risk assessment instrument that currently exists. Its greatest strength is however also its greatest weakness- this risk tool with unsurpassed validity is, unfortunately, unreal.

Oregon JCP FIRE Validation Dataset

The Oregon JCP FIRE Validation Dataset is a risk dataset that contains the records of 12,730 juveniles. However, to facilitate manageability, the original data are truncated to a smaller size of 1,000 cases using random selection. Its final size is significant in two ways. First, reducing the size of the data makes the manipulation and injection of error more tractable. Second, the master dataset "Risk Device X", to which the results of Oregon JCP FIRE are compared, also contains 1,000 cases. Matching the size of the two datasets makes the results more interpretable.

The risk device is comprised of 30 risk items, all of which or dichotomous (see table for distributions). The criterion variable that sets the base rate for this population is a variable designated as "follow-up referral." Generally, failure within this population is comprised of juveniles who are referred for follow-up action. The base rate for this population is .3, meaning that 30 percent of the juveniles fail within a follow-up period of 12 months. This risk device divides the population into four categories of risk: low, low/moderate, moderate, and high. The failure rate for each group is 16.7%, 28.9%, 37.27%, and 47.7% respectively. The scale score ranges from 0 to 28. The full description of the Oregon JCP FIRE Validation dataset can be found in the following table.

Table 7: *Scale Score for Oregon JCP Fire Risk Instrument*

Scale Score	N	Number of Failures	Number in Cutoff	Cutoff Base Rate (%)	Failure Rate
0	705	73			
1	989	120		Low	
2	1124	204	4872	788	16.17%
3	1094	205	(39.39%)	(22.41%)	
4	960	186			
5	902	225			
6	817	221	3179	Low/Moderate	
7	755	225	(25.7%)	919	28.9%
8	705	248		(26.13%)	
9	627	221			
10	532	169			
11	474	182	2401	Moderate	
12	393	159	(19.41%)	895	37.27%
13	375	164		(25.45%)	
14	365	155			
15	328	158			
16	285	142			
17	219	101			
18	172	72			
19	163	84			
20	117	61			
21	96	51	1918	High	47.7%
22	63	30	(15.51%)	915	
23	49	24		(26.02%)	
24	30	17			
25	14	10			
26	10	6			
27	6	4			
28	1	0			
Total	12370	3517			

Table 8: *Description of Oregon JCP Fire Instrument*

Item	Risk Item	Categories	Distributions (Percent)	
1	School attach	0,1	7751 (62.66%)	4619 (37.34%)
2	Truancy	0,1	9291 (75.11%)	3079 (24.88%)
3	Academic fail	0,1	7702 (62.26%)	4668 (37.74%)
4	Drop out	0,1	10991 (88.85%)	1379 (11.15%)
5	Friends bad behave	0,1	6119 (49.47%)	6251 (50.53%)
6	Friends drop out	0,1	5908 (47.76%)	6462 (52.24%)
7	Friends disapprove	0,1	7446 (60.19%)	4924 (39.81%)
8	Friends good academic	0,1	9973 (80.62%)	2397 (19.38%)
9	Adults friend	0,1	10654 (86.13%)	1716 (13.87%)
10	Behave before 13	0,1	9718 (78.65%)	2652 (21.44%)
11	Behave last month	0,1	10655 (86.14%)	1715 (13.86%)
12	Crim refs 3	0,1	9903 (80.06%)	2467 (19.94%)
13	Constructive school act	0,1	5455 (44.10%)	6915 (55.90%)
14	Runaway	0,1	11116 (89.86%)	1254 (10.14%)
15	Runaway recent	0,1	11405 (92.20%)	965 (7.8%)
16	Behave hurts others recent	0,1	10429 (84.31%)	1941 (15.69%)
17	Behave hurts self	0,1	8964 (72.47%)	3406 (27.53%)
18	Impulse aggression	0,1	9676 (78.22%)	2694 (21.78%)
19	Harms animals	0,1	12164 (98.33%)	206 (1.67%)
20	Weapons	0,1	11687 (94.43%)	689 (5.57%)
21	Fam communication	0,1	8985 (72.64%)	3385 (27.36%)
22	Poor fam supervision	0,1	8853 (71.57%)	3517 (28.43%)
23	Fam conflict	0,1	9289 (75.09%)	3081 (24.91%)
24	Cps cv	0,1	9390 (75.91%)	2980 (24.09%)
25	Crim family member	0,1	9675 (78.21%)	2695 (21.79%)
26	Sub use	0,1	8894 (71.90%)	3476 (28.1%)
27	Sub use prob	0,1	9240 (74.40%)	3130 (25.3%)
28	Sub use 13	0,1	10021 (81.01%)	2349 (18.99%)
29	Sub use school	0,1	10713 (86.60%)	1657 (13.40%)
30	Anti social	0,1	9680 (78.25%)	2690 (21.75%)

Critique

The Oregon JCP FIRE makes up an important part of the analytical process. Because it is a real dataset, it provides the researcher with a realistic glimpse of the relationships between the many elements in a risk dataset and instrument. For example, it

is evident that the Oregon JCP FIRE, which Baird et al (2012) claims to be one of the best juvenile risk assessment instruments, is not a better risk assessment instrument then then Risk Device X. Another interesting relationship found in the Oregon JCP FIRE is that it contains highly intercorrelated risk items. Nonetheless, the use of the Oregon JCP FIRE to test the hypotheses, where possible, would lend great support to findings found from Risk Device X.

The primary reason that the Oregon JCP FIRE could not be used as the sole dataset from which hypotheses tests are made is because it lacks the variety needed to test the different hypotheses. For instance, such risk instrument contains 30 dichotomous risk variables that are largely uniform, of which most are positively skewed. It, thereby, does not offer the opportunity to compare the sensitivity of error impacted by different types of risk items, such as those with three or more categories or those that are negatively or normally skewed. Thus, the inherent design of the Oregon JCP FIRE does not make it a good fit to answer most of the current study's research questions.

Oregon JCP FIRE 3 Categories

To help test several hypotheses, the risk categories in Oregon JCP FIRE was reduced from the original four categories to three. In every aspect, this instrument and dataset are similar to the original Oregon JCP FIRE. The rationale behind shrinking the number of risk categories for this risk tool comes from the arguments made by Baird (2009) and Todd Clear (personal communications, 2013). Having three categories of risk is more beneficial under most circumstance because: 1) most agencies do not match risk tools with sufficient levels of supervision, and 2) most levels of supervision are not

meaningfully different to differentially impact an individual's actual likelihood of reoffending. As well, matching the number of risk categories across the risk tools facilitates the comparison of results. The distribution and scale score for this instrument is shown below.

Table 9: *Scale score of Oregon JCP FIRE with 3 Categories*

Scale Score	N	Number of Failures	Cases in Cutoff	Cutoff Base-Rate (%)
0	70	6		
1	104	14	409	69 (16.8%)
2	78	16		
3	84	15		
4	73	19		
5	70	16		
6	60	18		
7	63	23	293	81 (27.6%)
8	49	15		
9	51	19		
10	38	12		
11	33	13		
12	34	16		
13	30	12		
14	37	21		
15	25	14		
16	22	11		
17	23	10	298	179 (50%)
18	12	8		
19	16	4		
20	9	5		
21	4	1		
22	7	4		
23	3	2		
24	3	3		
26	2	2		
Total	1,000	299		

Oregon JCP FIRE 3 Categories- 11 items (Burgess Method)

This version of the JCP FIRE is made up of 11 items, as opposed to the 30 items in the original instrument. To create this instrument, the same data from the Oregon JCP FIRE were entered into a logistic regression, and only 11 of the most robust variables are

retained. Several researchers have cited the importance of parsimony (Austin et al., 2003; Flores et al., 2004). Eliminating excesses in risk items significantly augmented validity. In addition, this instrument codes each of the 11 risk item using the Burgess Method, where each item dichotomy is coded as 0 or 1 to form the total risk score.

Table 10: *Scale score for Oregon JCP-Burgess*

Scale Score	N	Number of Failures	Cases in Cutoff	Cutoff Base-Rate (%)
0	215	34		
1	211	52	426	86 (20.1%)
2	198	62		
3	145	52	343	114 (33.2%)
4	95	41		
5	64	21		
6	38	19		
7	19	8	232	100 (43.1%)
8	10	8		
9	3	2		
10	3	1		
Total	1,000	300		

Oregon JCP FIRE 3 Categories – 11 items (Coefficients)

This version of the JCP FIRE is also made up of 11 risk items. It is different from the "Oregon JCP FIRE -11 items (Burgess Method) because it scores individuals by the actual weight contribution of each risk item. After adding the 11 risk items into a logistic regression with the outcome variable, it is learned that each item varies greatly on its individual contribution to the risk function. Thus, in this version of the Oregon JCP FIRE, the individuals are scored by the actual weights or coefficients of each item. As a result of scoring the variables by their individual coefficients, the range of scale score expanded enormously, from 0 to 35.

Table 11: *Scale score for Oregon JCP-Coefficients*

Scale Score	N	Number of Failures	Cases in Cutoff	Cutoff Base-Rate (%)
0	215	34		
2	85	16	394	77 (19.5%)
3	94	27		
4	25	6		
5	105	31		
6	37	12		
7	35	6	351	107 (30.4%)
8	72	25		
9	21	12		
10	56	15		
11	42	20		
12	18	7		
13	45	16		
14	15	8		
15	21	12		
16	27	11		
17	4	1		
18	13	5		
19	14	8		
20	5	1		
21	13	6	255	116 (45.4%)
22	4	2		
23	8	1		
24	9	6		
25	3	2		
26	1	1		
27	7	5		
28	1	1		
30	2	1		
33	1	1		
35	3	1		
Total	1,001	300		

Risk Device X- Version 5 Categories

To understand how differences in the number of risk categories impact the sensitivity of error, another version of Risk Device X with 5 risk categories is created. Here, "Risk Device X Version 5 Categories" is an exact copy of the original Risk Device X. The only difference is that the individuals are divided into 5 risk groups as opposed to 3. The construction of this instrument is accomplished by recoding the original scores into ones that divide the offenders into more groups. The following table explains the scale scores and different properties found in the new version of Risk Device X.

Table 12: *Scale Score for Risk Device X- Version 5 Categories*

Scale Score	N	Number of Failures	Number in Cutoff	Failure rate (%)
7	2	0		
8	2	0		
9	3	0		
10	14	1	183 (18.3%)	(2.5%)
11	31	2		
12	57	2		
13	74	5		
14	93	10		
15	102	12	195 (19.5%)	(5.5%)
16	119	12		
17	109	22		(8.5%)
18	105	75	228 (22.8%)	(39.1%)
19	90	80		
20	74	64		
21	43	37	195 (19.5%)	
22	39	36		
23	23	21		(44.3%)
24	13	11		
25	6	6	199 (19.9%)	
26	0	0		
27	1	1		
Total	1000	397	1000	

Procedure for Injecting Error

The next step involves the injection of error into the dataset. There are two types of error of interest in the current study: random error and systematic error. Random error conveys a sense of randomness in how the numbers are corrupted. In other words, an individual could belong to a higher or lower risk option if he is incorrectly placed in a medium risk category for a risk item. Systematic error, on the other hand, constitutes incorrect placements that are consistently made towards a specific direction.

The difference between random and systematic error have deep implications for how error is defined and created in the study. The study increases the level of error by intervals of 10 (ie. 10 percent, 20 percent, and 30 percent) to understand how different quantities of error impact the risk items. Before assigning error to the cases, one must identify the cases to which error is assigned. This can be accomplished by using the random number generator to assign numbers to the 1,000 cases. If 10 percent error is desired, then 100 cases for each risk item with the lowest number (randomly created) can be selected for the injection of error. For each risk item, randomization is used to select the cases. As such, it is possible for the same individual to experience changes in more than one risk item. As well, it is possible that portions of individuals will never be subject to receiving any injected error. The same method can be repeated to test the impact of injecting 20 percent or 30 percent error.

Random error is replicated by shifting a case to a different category. For example, if a case for a risk item is assigned to one of three risk categories (i.e. low, medium, high), error would mean that the case actually belongs to one of the other two categories. And

if 10 percent error is expected for one risk item, then 10 percent of the cases are expect to belong to one of the other two options instead. On the other hand, systematic error is replicated by shifting cases towards a desired direction. For example, 10 percent systematic error towards the direction of underestimation translates into a shift of 10 percent of the cases towards a higher risk category. This same general procedure for injecting error should be repeated for each risk item.

Research Hypothesis

The primary research question is: how do different classification instrument properties affect the transfer of error. This research inquiry can be answered by focusing on more specific inquiries outlined below. Each research question can be addressed by testing its respective hypotheses.

1) *What is the impact of error on risk assessment instruments? H1-H4*
2) *How do different distributions of cases across categories impact sensitivity? H5-H6*
3) *How do different distributions of cases across risk variables affect the sensitivity of error? H7-H11*
4) *Does the expansion of scale scores in risk instruments increase the sensitivity of error? H12*

Based on these research questions, the proposed study examines the sensitivity of error in offender risk assessment instruments by using both conceptual data and actual data. After exploring the effects of random error and systematic error on four risk devices, the study will examine the impact of different distributions of cases across

categories on sensitivity. Next, the impact of distributions of cases across risk variables on sensitivity is examined. Finally, the study will explore the interplay between range of scores and sensitivity. The following hypotheses are proposed:

H1: After injecting random error into entire risk models, the level of classification error is equal to the level of error that is injected into the risk items.

After injecting different levels of random error into the risk items, final classification of offenders should experience a level of error that is commensurate of the percentage of error that was initially injected into risk items. The effect of random error is impartial to any particular risk category. It affects all segments of a risk item equally. Thus, if we assume that there is a 10 percent error in the variables, we could also assume that the error is split evenly among all the possible outcomes. In the final classification phase, an equal amount of classification error should be seen.

H2: After injecting systematic error (up and down) into entire risk models, the level of classification error is equal to the level of error that is injected into the risk items.

Since the injection of systematic error effects unidirectional changes in the scores, the magnitude of misclassification should be greater. On the other hand, random error creates changes that could either increase or decrease a risk score, moderating the magnitude of misclassification.

H3: After injecting random error into each risk model, the subgroup base-rates for high-risk groups will decrease and the subgroup base-rates for low risk groups will increase, showing a pattern of regression towards the mean in such rates.

It is logical to assume that error will reduce the instrument's ability to classify individuals. As such, the injection of error should cause subgroup base-rates from high and low categories to shrinks towards the mean. As the shrinkage of the subgroup base-rates continues, the validity or the capacity of the risk instrument to divide individuals into groups with meaningfully different recidivism rates would be compromised.

H4: After injecting systematic (up and down) error into each risk model, the subgroup base-rates for high-risk groups will decrease and the subgroup base-rates for low risk groups will increase, showing a pattern of regression towards the mean in such rates.

H5: Risk instruments that over-classify high risk individuals will be more sensitive to random error.

Risk instruments, depending on a myriad of factors, can distribute the majority of the cases to several risk categories, causing disproportionate clusters of cases in certain risk groups. As such, different skews can be formed. It is argued that negatively skewed cases in the outcomes of the entire risk classification instrument will have the greatest impact on increasing sensitivity. Here, it is important not to confuse the individual distribution of risk items (mentioned in hypothesis #7) with the distribution of the collective classification of cases.

H6: *Increasing the number of categories to which offenders are classified will increase misclassification and the sensitivity of error.*

Increasing the number of risk items in risk devices will reduce the tendency for misclassification error, thereby reducing sensitivity. The logic behind such supposition

comes from the natural expansion of scores that will comprise a function, ensuing the addition of risk items. The expansion of scores will increase the range of scores that form each risk category. Thus, risk instruments with more risk variables are prone to have less misclassification.

H7: Negatively skewed risk items will have the greatest impact on increasing the sensitivity of error, when compared to normally and positively skewed risk items.

Of the three general types of distributions that can exist in the scores for each case, negative skews are speculated to have the gravest impact on misclassification. To fully understand the implications of distributions, the trajectory of error for each distribution type requires individual analysis and explanation.

Normal Distributions: Error in normally distributed items will push the majority of cases to the outer ends or into more extreme risk categories. Normal distributions are symmetrical and usually experience a cluster of cases in the middle categories (Field, 2005). If error is injected into such distributions, the effect will be that most of the error will be experienced by the category under which most cases will fall. Thus, if all the risk times in an offender risk assessment instrument are normally distributed, most of the error in the final classification phase will be in the medium risk categories. Logically then, the error will cause the cases to displace outwards to the more extreme categories.

Skewness: is often associated with the lack of symmetry in frequency distributions. If it is positively or negatively skewed, the majority of cases is clustered on either end, thereby yielding more errors in these extremes. Thus, the skewness of a

distribution should significantly alter final classification outcomes if random error is injected.

Positive skew: When random error is combined with positively skewed distributions in risk categories, the direction of impact will be towards higher risk categories. A positive skew clusters cases on the right side of the distribution and tails off to the right (Field, 2005). Thus, random error will more likely affect these categories and reduce the skew of the distribution; that is, random error will push the cases to the right or higher risk categories during final risk classification.

Negative Skew: The injection of random error into negatively skewed distributions in risk categories will have the gravest impact on misclassification. A negative skew illustrates that a cluster of cases are on the right side of the distribution. In this situation, random error will likely push the cases towards the left or the lower risk categories. The final classification outcome will likely be significantly pushed towards the left, falsely identifying higher amounts of low risk offenders.

The assumption that negatively skewed items will have the greatest impact on sensitivity is primarily warranted by the understanding that the subgroup base-rate progressively increases from low leveled to higher leveled risk categories. In other words, a valid risk device should show a positive relationship between risk and number of failures within the subgroups. And for this reason, anytime a significant amount of cases move from a higher-level category to a lower level category, the sub-group base rate will increase in lower level categories. When sufficient changes to the subgroup base-rate are induced so that subgroup base rates are no longer commensurate with risk, the entire risk

instrument would be rendered invalid. Negative skews would invariably provoke this outcome and displace failed cases to lower leveled risk categories.

Though the magnitude of displacement should not vary from one type of distribution to another, error in negatively skewed items will have the most significant impact on the validity of the risk device. This is because most other distributions only superficially move cases around, while error in negative distributions shifts cases that directly impact the base rates.

Error in dichotomous variables is unidirectional. If it doesn't belong to the existing category, then it belongs to the other category. The destination to which cases are displaced is absolute. However, the injection of error into risk items with more categories has more places to go. Thus, error in dichotomous variables displaces cases more aggressively because the shift of all of the erroneous cases can only go towards one direction.

H8: *Increasing the peakness of distributions in risk items will increase the amount of error or misclassifications towards a particular direction*.

According to Fields (2005), the kurtosis of a distribution refers to the pointyness of a distribution. In terms of kurtosis, a distribution can either be leptokurtic or platykurtic. If it is normal or average in peakness, it is called mesokurtic. Leptokurtic distributions are more peaked, whereas platykurtic distributions are flatter. It is hypothesized that the direction of misclassified cases is largely determined by the variable's kurtosis. For instance, the injection of error into flatly distributed variables should mean that misclassification can go in any direction. However, as the peakness

increases, more cases are moving towards one direction, thereby dictating the magnitude of such shifts.

H9: Dichotomous risk items will have a greater impact on the sensitivity of error than will risk items with more category options.

Today, most risk tools are comprised of mostly dichotomies for the purpose of simplicity. It is the contention of this study that dichotomies, as compared to risk variables with three or four categories, worsen the problem of misclassification.

H10: Increasing the weight of individual risk items will increase the sensitivity of error.

Risk items can vary greatly in terms of the score points they individually contribute to the overall risk function. Though some instruments, for the purpose of increased manageability, are strictly designed so that every risk item contributes the same score to the entire function, other instruments may contain an array of risk items that would differentially contribute to the risk function. It is hypothesized that the injection of error into such risk items, whose score-point contribution is greater than the score-point contribution of its counterparts, will have a greater impact on misclassification.

H11: Random error will produce less misclassification than will systematic error.

This hypothesis will examine the impact of random and systematic error on individual risk items. It is hypothesized that random error, which is directionless, will produce lower levels of misclassifications. Conversely, systematic error will increase misclassifications because it pushes error towards the same direction.

H12: Adding more risk items to a risk tool will decrease the sensitivity of error.

Offender risk classification instruments could vary greatly on the number of risk items they contain. While some risk designers have argued for the need of longer and more comprehensive instruments, some have argued for parsimony. To reconcile this debate, the issue about the appropriate number of risk items that should comprise a risk tool is analyzed from a sensitivity perspective. It is hypothesized that increasing the number of risk variables would subsequently expand an instrument's range of score, and the range of score will provide a buffer for error.

After understanding the impact of error on risk tools, the next objective is to identify risk properties that would increase the sensitivity of error. It is hoped that by increasing the understanding of the interplay between various risk properties and the sensitivity of error, more efficient risk tools could be constructed.

Analytical Plan

Each hypothesis is specifically designed to answer its respective research question. Hence, the hypotheses are different and require different types of analyses to answer them. The course of action for answering the hypotheses is outlined below.

Hypothesis #1 & #2:

After random error is injected into the four risk tools, the misclassifications are calculated. To answer this hypothesis, the original error that is injected into each instrument is compared to the level of misclassification that yields. For example, if 10 percent of cases are misclassified following the injection of 10 percent error, then the

levels of error are considered equal. For hypothesis #2, the general procedure is followed. After systematic (up and down) errors are injected into the four risk tools, the misclassifications are calculated and compared to the systematic error that is injected into the risk data.

Hypothesis #3 & #4:

To understand the change in subgroup base-rates, the new subgroup base-rates are subtracted from the original subgroup base-rates. The level of change is, thus, measured by the magnitude of such differences. If the direction of change for low and high risk groups is towards the middle, then it fails to reject hypotheses #3 and #4.

Hypothesis #5:

To test hypothesis #5, new cutoffs are drawn for Risk Device X. Three versions of Risk Device X are constructed, each with a tendency to over-classify individuals into low, moderate, or high risk. The process of injecting error into each version is implemented. The number of misclassified cases is calculated for each version of Risk Device X to see how over-classification of individuals into specific categories would impact sensitivity.

Hypothesis #6:

Misclassifications produced from the injection of error into Risk Device X and Risk Device X-5 Cats are compared. Risk Device X-5 Cats is different because it contains cutoffs that divide individuals into five categories of risk.

Hypothesis #7:

Risk Device X contains variables with different skews. Variables A, G, and H are negatively skewed. Variables B, F, J, and K are positively skewed. And Variables C, D, E, and I are normally skewed. Misclassifications following the injection of error are calculated for each individual item to understand the relationship between skew and sensitivity.

Hypothesis #8:

Oregon JCP FIRE is used to test this hypothesis. To ensure that the levels of change is not a function of probability, 5 pointiest risk items with a positive skew are selected. As well 5 flattest risk items with a positive are selected. After injecting random error into Oregon JCP FIRE, two types of changes are expected, misclassifications toward a higher risk level and misclassifications toward a lower risk level. The total misclassifications going towards each direction (high and low) are then calculated to form two averages. A ratio is then formed by comparing the cases that are misclassified to a higher risk level to cases that are misclassified to a lower risk level. The greater the disproportionality in this ratio, the greater the magnitude of misclassification towards a direction is indicated.

Hypothesis #9:

Items E, F, and G from Risk Device X are dichotomous. The misclassifications resulting from the injection of error into such risk items can be compared to the ones resulting from the injection of error into the other risk items.

Hypothesis #10:

Items C and J from Risk Device X are weighed disproportionately more than their counterparts. The misclassifications for these items can be compared to their counterparts to understand how the increase of weights contributes to sensitivity.

Hypothesis #11:

To test answer hypotheses #7, #8, #9, and #10, random error was injected into relevant risk items. For hypothesis #11, the same analyses can be done. But this time, systematic error is injected into the items instead.

Hypothesis #12:

Misclassifications for Oregon JCP FIRE and Oregon JCP-Burgess are compared. Since Oregon JCP-Burgess is simply a truncated version of Oregon JCP FIRE, a simple comparison of classification error for each risk tool would be adequate to answer hypothesis #12.

Analytical Procedure

Finally, the analysis segment of the study needs to be explained in further detail. Prior to injecting error into the cases, reasonable cutoffs that define the upper and lower limits for each category are set. In other words, the number of individuals who constitute each risk category needs to be determined immediately after the cases are randomly generated. The analysis is conducted primarily by comparing the group sizes of each risk category before the injection of error to the new group sizes for each risk group after the injection of error. Many statistical methods can be used to measure the differences in

group sizes. Cross tabs, in particular, are appropriate for this kind of analysis. The injection of error is expected to significantly alter the sizes of each risk group. The cross tabs test for significance in the differences. The specific procedure is as follows:

Step 1: Determine the risk classification designation for each individual by adding up their score from all 11 risk items.

Step 2: Once the classification of all the individuals are set, the injection of error can be implemented on the variables by following the procedure for injecting error.

Step 3: For each time that error is injected into a variable, the resulting shift in cases may or may not impact the risk classification designation for an individual.

Step 4: A simple comparison can be done between the risk classification designations of individuals before and after the injection of error. Some classification experts suggest that a simple comparison of group sizes of each risk category will give the clearest indication of validity. However, cross tabs can be used to give us a better idea of differences and their significance.

To cross check, one can compare the change in base rates resulting after the injection of error across the items. Relative to base-rate changes in other items, significant changes in a particular item will provide comparative value and a clear indication of heightened sensitivity within the item.

Comparing the group sizes for each risk category before and after the injection of error for a single risk item tells us very little about the impact of error on risk devices. However, we hope to recognize a pattern in the risk devices by comparing the before and

after effect of error for multiple risk item. By the end, it is hoped that the analysis will give us a clearer idea on what properties in risk devices increases the sensitivity of error.

The steps described above are for one single distribution. To understand how error affects other distributional scenarios, the steps need to be repeated for each statistical distribution in question. Also, a major premise of the study is to measure the tolerance for each statistical distribution. This means that different levels of error need to be injected into the distributions. This will allow the researcher to see change on a continuum.

Measurement

The current study relies on three specific methods to measure the impact of error on sensitivity. They are: 1) individual case displacements, 2) subgroup base-rate change, and 3) significance.

First, individual case displacements refer to the change in individual's risk designation subsequent the injection of error. The study is premised on the idea that increases or decreases in an individual's risk score does not necessarily translate into changes in the classification of risk. Thus, for example, the level of initial error should not be tantamount to classification error. The case shifts look at how a person's risk designation is changed by the injection of error. For example, an individual was classified as low risk when he should had been classified as moderate or high risk. Thus, a change in risk designation for an individual will count as one displacement.

Second, subgroup base –rate change refers to how the base rate is altered after the injection of error. This type of measurement directly taps into the core of study because

it examines the elements that make an instrument valid, the subgroup base rate. The subgroup base rate is calculated by dividing the N by the number of failures in its respective category. For example, if there were 100 individuals and 10 failures in the low risk category, then the subgroup base-rate or recidivism rate would be 10%. This measurement will examine the change in base rate for each risk category after error is injected, which translates to subtracting the original subgroup base rate with the new subgroup base-rate. Using the same scenario, for example, if the injection of error produces a new base rate of 14 percent, we would have a change of 4 percent in the base rate for the low risk group. And to standardize this measurement for an entire model, a net change calculation can be created, where the change in base-rates across the different categories of risk are summed.

Finally, significance tests are administered to the subgroup base-rate changes. The formula for the significance test is: $Z= (p-P)/ \text{sqrt}((PQ/N))$, where p = failure rate for item with error, P = true failure rate, Q = 1 - true failure rate, and N = number of cases. Though this significance test empowers the study because it provides a popularly recognized and standardized way to measure the impact of error, it needs to be interpreted with some level of caution, especially when we are comparing the validity of entire models. For example, if there are three risk categories that which constitute a classification instrument, having significance in the changes of base-rates for all three categories may not necessarily mean that the model is less tolerant of error, though generally it would be so suggested. An unforeseen circumstance might have emerged where the majority of change is heavily clustered in one category. Thus, the

interpretation of significance needs to be complemented with the equally important interpretation of effect sizes.

These three methods by which the sensitivity of error is measured are equally important. Each is adept and appropriate to answer different research questions, depending on the specific circumstances.

Chapter 5- Assessing the Sensitivity of Error in Risk Devices- Analyses and Results

This chapter assesses the hypotheses described in the previous chapter to determine the relationship between different risk properties and the sensitivity of error in order to address the proposed research questions. This chapter is organized as follows. The first section examines the overall impact of initial error, both random and systematic, on classification error. The second section describes the results that looked at the distribution of cases across risk categories. In the third section, the impact of distributions of cases for individual variables on sensitivity is discussed. Finally, the last section will look at the relationship between the range of score and sensitivity, as well as, looking at specific circumstances where the range of score is augmented.

Research Question #1: What is the impact of error on misclassifications in Risk Device X?

The first phase of this research explores the overall transfer of error in offender risk classification instruments. That is, it examines the relationship between errors in risk information and errors in classification outcomes. A primary reason pushing forth the current study has to do with the paucity of knowledge about the impact of error. As such, the sensitivity of error and classification effectiveness of risk tools has been critically understudied. It is generally assumed that if we were to inject an identifiable quantity of error in the risk items, a similar proportion of error will take place in the classification outcomes. However, such supposition has not been empirically supported. The answer to this question is, unfortunately, complicated as well.

Research Question #1 can be broken down into two segments. The first section concerns the number of misclassified cases that yields after the injection of 10 percent

random error, 20 percent random error, 10 percent systematic (up) error, and 10 percent systematic (down) error. Section two will examine the impact of such errors by looking at the changes in subgroup recidivism/base rates.

Results- Hypothesis #1:

- After injecting random error into entire risk models, the level of classification error is equal to the level of error that is injected into the risk items.

The study rejects hypothesis #1, which states that the level of error injected into the risk items is equal to the level of misclassifications that occur. Table 1A shows that 10 percent random error does not cause an equal amount of classification error. Instead, 10 percent random error in the risk items causes 24.4% misclassification in Risk Device X, 26.3% misclassification in Oregon JCP-3 Cat, 33.7% misclassification in Oregon JCP-Burgess, and 35.2% misclassification in Oregon JCP-Coefficients (see table 1A). Thus, small levels of error across the risk items in a risk tool yield high levels of classification error. Similar results are found after injecting 20 percent random error into the same risk tools. Twenty percent random error causes 35.9% misclassifications in Risk Device X, 48.3% misclassifications in Oregon JCP-3 Cats, 52.7% misclassifications in Oregon JCP-Burgess, and 53.7% misclassifications in Oregon JCP-Coefficients (see table 1B). Hence, random error does not cause equal levels of classification error, it causes disproportionately more. Such exorbitant levels of classification error that yield, warrant much doubt and questioning about the criminal justice system's over dependence on classification systems.

Next, it is visible that different risk tools are consistently more tolerant of error. Of the 4 risk instruments tested in the study, Risk Device X fared the best in tolerating error, while Oregon JCP-Coefficients fared the worst. That is, injecting 10 percent random error into its risk items caused 24.4% misclassification in Risk Device X, while causing 35.2% misclassifications in Oregon JCP-Coefficients. Compared to these instruments, Oregon JCP FIRE and Oregon JCP-Burgess are in the middle in terms of demonstrating the capacity to tolerate error.

Comparing Burgess method and Coefficient method

There are two popular methods of calculating scores in a risk model. The Burgess method is boasted for its simplicity in determining offender risk scores. Dichotomies are coded either 0 for not displaying a risk trait or 1 for displaying a risk trait. The total score is calculated by summing such individual scores. The Coefficient Method utilizes the weights of individual variables that are obtained from regressions to calculate offender risk. Each of these methods has a profound impact on an instrument's range of score. The Burgess method creates smaller scales scores, while the Coefficient Method creates vaster scale scores.

In terms of reducing the sensitivity of error, the study suggests that the Burgess Method of risk score construction is superior to the Coefficient Method. For instance, the injection of 10 percent random error produces 337 misclassified cases in Oregon JCP-Burgess and 352 misclassified cases in Oregon JCP-Coefficients. Similarly, 20 percent random error produces 527 misclassified cases in Oregon JCP-Burgess and 537 misclassified cases in Oregon JCP-Coefficients. Thus, Oregon JCP-Burgess, which was

constructed using the Burgess method, has a lower tendency for misclassification after the injection of random error

Random Error and Direction of Misclassifications

Random error causes misclassifications of cases to go in both directions. In other words, fair amounts of cases are displaced towards higher risk levels, while fair amounts of cases are displaced down towards lower risk levels. Thus, random error causes random misclassifications, or cases to displace up (towards high risk) and down (towards low risk). For instance, of the 244 cases that are displaced in Risk Device X, 133 cases moved to a higher risk level and 111 moved to a lower risk level. The amount of cases that displaced upwards is roughly equal to the amount of cases that displaced downwards.

However, equal proportions of misclassifications are not seen for the other risk tools, i.e. Oregon JCP FIRE, Oregon JCP-Burgess, and Oregon JCP-Coefficients. For such risk devices, random error causes many more misclassified cases to displace towards higher risk levels, instead of displacing towards lower risk levels. Take Oregon JCP FIRE for instance. Of the 263 cases that are misclassified after injecting 10 percent error, a significant portion of those moved up (245 cases), compared to those that moved down (18 cases). Similar disproportions in misclassified cases are seen for Oregon JCP-Burgess and Oregon JCP-Coefficients.

Proportionality- Error in risk information and error in classification outcomes

Increases in the level of error that is injected into the risk information also increase the level of misclassification error. Across all four risk tools, similar patterns of displacement are experienced after increasing random error from 10 percent to 20 percent,

suggestive of proportionality. For instance, the injection of 10 percent error into the instrument Oregon JCP FIRE causes 26.3% of the cases to be misclassified (see Table 1A). Injecting 20 percent random error into the same risk tool caused 48.3% of the cases to be misclassified (see Table 1B). Thus, the overall number of misclassified cases increases when the level of error in the risk information is increased. Similar patterns of change are seen for other risk tools. Increasing the level of error from 10 to 20 percent produced: 115 more misclassified cases in Risk Device X (359- 244); 190 more misclassified cases in Oregon JCP FIRE- Version 11 Items: Burgess Method (527-337); and 185 more misclassified cases in Oregon JCP FIRE- Version 11 Items: Coefficients (537-352). Hence, increasing the initial error by two-folds, from 10 percent to 20 percent, creates a proportionate increase in misclassification error in all four risk tools, which generally shows a net increase of two-folds.

Table 1A: *Number of Misclassifications after Injection of 10% Random Error*

Risk Tool	Low to Moderate	Moderate to High	Low to High	High to Moderate	Moderate to Low	High to Low	Total
Risk Device X	56	64	13	57	50	4	244
Oregon JCP FIRE (3 cats)	187	58	0	14	4	0	263
Oregon JCP-Burgess	187	92	13	28	17	0	337
Oregon JCP Coefficients	164	99	43	29	17	0	352

Table 1B: *Number of Misclassifications after Injecting 20% Random Error*

Risk Tool	Low to Moderate	Moderate to High	Low to High	High to Moderate	Moderate to Low	High to Low	Total (N= 1000)
Risk Device X	85	109	26	65	62	12	359
Oregon JCP FIRE (3 cats)	333	116	18	13	3	0	483
Oregon JCP-Burgess	211	171	103	27	14	1	527
Oregon JCP Coefficients	176	194	128	26	13	0	537

Results- Hypothesis #2:

- After injecting systematic error (up and down) into entire risk models, the level of classification error is equal to the level of error that is injected into the risk items.

Systematic Upward Error

The injection of systematic (up) error into the risk instruments has resulted in dissimilar patterns of misclassifications. Hypothesis #2 is rejected. The quantity of misclassifications is not equal to the level of systematic error that was injection into the risk tools. Specifically, 10 percent systematic (up) error causes 18.4% misclassifications in Risk Device X, 29.8% misclassifications in Oregon JCP FIRE (3 cats), 33.2% misclassifications in Oregon JCP-Burgess, and 33.1% misclassifications in Oregon JCP-Coefficients (see table 2A). In all four of the risk tools, low levels of error causes great levels of classification error.

Different risk tools have different levels of tolerance for systematic (up) error. Risk Device X fared the best, while the Oregon JCP-Burgess fared the worst. Oregon JCP-Burgess misclassifies individuals 18% more than Risk Device X. Oregon JCP and Oregon JCP-Coefficients are in the middle in terms of their ability to tolerate systematic (up) error.

Systematic (up) error causes the misclassification of cases to go to a higher risk level. Hence, none of the cases are displaced to lower risk levels after the injection of such error. Table 1C shows the impact of systematic (up) error across the four risk tools.

Table 2A: *Number of Misclassifications after Injection of 10% Systematic (Up) Error*

Risk Tool	Low to Moderate	Moderate to High	Low to High	High to Moderate	Moderate to Low	High to Low	Total (N= 1000)
Risk Device X	73	104	7	0	0	0	184
Oregon JCP FIRE (3 cats)	206	92	0	0	0	0	298
Oregon JCP-Burgess	197	120	15	0	0	0	332
Oregon JCP Coefficients	172	115	44	0	0	0	331

Table 2B: *Number of Misclassifications after Injection of 10% Systematic (Down) Error*

Risk Tool	Low to Moderate	Moderate to High	Low to High	High to Moderate	Moderate to Low	High to Low	Total (N= 1000)
Risk Device X	0	0	0	79	70	2	151
Oregon JCP FIRE (3 cats)	0	0	0	41	39	0	80
Oregon JCP-Burgess	0	0	0	41	44	1	86
Oregon JCP Coefficients	0	0	0	40	41	1	82

Systematic Downward Error

The study rejects hypothesis #2. Systematic downward error does not produce equal levels of misclassifications. The injection of 10 percent systematic downward error causes 15.1% of the original cases to be misclassified in Risk Device X, 8% in Oregon JCP FIRE, 8.6% in Oregon JCP-Burgess, and 8.2% in Oregon JCP-Coefficients (see table 2B). Of the four risk tools, Oregon JCP FIRE has the biggest capacity to tolerate error, followed by Oregon JCP-Coefficients, Oregon JCP-Burgess, and Risk Device X. Systematic downward error, in all three Oregon JCP instruments, causes lower levels of misclassifications. Finally, systematic downward error also causes the misclassification of cases to go in one direction, towards lower risk levels.

Results- Hypothesis #3:

- After injecting random error into each risk model, the subgroup base-rates for high risk groups will decrease and the subgroup base-rates for low risk groups will increase, showing a pattern of regression towards the mean in such rates.

Hypothesis #3 states that increases in initial error will cause recidivism rates to regress towards the mean. In other words, the injection of error will cause the recidivism rate in high-risk categories to shrink towards the average recidivism rate, cause the recidivism rate in low risk group to increase, and cause little change for the recidivism rate for mid-range risk categories. As was mentioned in chapter one, the measurement of sensitivity using subgroup base rates is tremendously useful. Clear (1988) suggests that validity should be measured by how well an instrument divides a population based on its recidivism rate. Thus, the summation of total shifts for each risk tool, as shown in Table

1A and 1B, communicates about the quantity of classification error, which is not a particularly useful measure to answer about a risk tools overall validity.

On the other hand, the changes in subgroup base rates, as shown in Table 3A speak to the significance of the shifts, and whether validity for the instrument has been compromised. In Table 3A, it is clear that the base rate for each risk group gravitates towards the mean after random error is injected into the risk instruments. For example, Risk Device X, JCP Burgess, and JCP Coefficients, all experienced movements in their recidivism rate that is towards the mean, which can be interpreted as the diminution of the level of validity for these risk tools. When recidivism rates move towards the middle, such rates become numerically closer to each other. As such, the risk tools lose the power to discern individuals based on their propensity for future reoffending. The result is an invalid risk instrument. The best-case scenario, obviously, would be the divergence of such recidivism rates, indicating increases in validity.

The interpretation of the changes in subgroup base-rates is complicated as well. There is some difficulty in comparing such changes across different risk tools. Unless the initial subgroup base rates are nearly identical for every category, a direct comparison would not be accurate. For instance, the changes experienced by both risk tools, Risk Device X and Oregon JCP FIRE, could not be directly compared due to their stark disparity in the original subgroup base rates (see Table 3A). Risk Device X is a better risk instrument by far; its ability to separate individuals into risk groups with meaningfully different subgroup recidivism rates is exceedingly superior to the one of Oregon JCP FIRE. In particular, for Risk Device X, the contrast between the recidivism rates for the high risk (84%) and low risk (7.4%) groups is roughly 78%, while the

contrast between the recidivism rates for the high risk (49.7%) and low risk groups (17.1%) in the Oregon JCP FIRE is roughly 32% (see Table 3A). This initial disparity makes a simple comparison between risk tools with varying levels of subgroup recidivism rates a difficult task.

The findings fail to reject hypothesis #3. The injection of error into the risk tools causes the convergence of subgroup base-rates, and increases to the level of such error (i.e. 10% to 20%) cause greater levels of convergences. In other words, the subgroup base rates will progressively shrink towards the mean as the level of random error increases, thereby reducing the validity of the risk tools. After injecting different levels of random error, such shrinkage is seen for three of the four risk tools (Risk Device X, Oregon JCP-Burgess, Oregon JCP-Coefficients). For example, the original low (20.1%), moderate (33.2%), and high risk (43.1%) groups in Oregon JCP-Burgess became 21.8%, 28%, and 39.1% following the injection of 10 percent random error, and 22%, 25.2%, and 45.9% following the injection of 20 percent random error, respectively. In each of these risk tools, the injection of random error caused the high-risk recidivism rate to drop while causing the low-risk recidivism rate to go up (see Table 3A).

The subgroup base-rates for one risk tool, however, did exhibit a different type of change in the subgroup base-rates. Following the injection of 10 and 20 percent random error into Oregon JCP FIRE, the low-risk group's recidivism rate diverged away from the mean, instead of moving towards it. The change in the recidivism rate for the high-risk group is consistent with those in the other risk tools. The high-risk group's recidivism rate change is towards the mean, similar to those changes experienced in other risk tools. For example, the original low (17.1%), moderate (31.8%), and high risk (47.9) groups in

Oregon JCP FIRE became 15%, 29%, and 44% following the injection of 10 percent random error, and 14.7%, 23.8%, and 42.8% following the injection of 20 percent random error, respectively (see table 3A). Thus, some more investigation of the Oregon JCP FIRE is warranted because it has shown that the injection of both 10 and 20 percent error creates higher levels of validity. This aberration in the change of subgroup recidivism rates is speculated to be either caused by randomness and chance or the uniquely uniform skews of the risk items.

Table 3A: *Subgroup Base-rate Changes after Injecting Random Error in Four Risk Tools*

Risk Tool	Level of Random Error	Low (Base-rate)	Moderate	High
Risk Device X	None	7.2% (20)	13.9% (46)	84% (331)
	After 10 Percent Random Error	10.8% (30)	22.1% (70)	71.7% (294)
	After 20 Percent Random Error	11.2% (27)	25.5% (79)	64.3% (291)
Oregon JCP FIRE (3 CATS)	None	17.1% (70)	31.8% (116)	49.7% (113)
	After 10 Percent Random Error	15% (34)	29% (146)	44% (119)
	After 20 Percent Random Error	14.7% (9)	23.8% (141)	42.8% (149)
Oregon JCP FIRE-Version (11 Items-Burgess Method)	None	20.1% (86)	33.2% (114)	43.1% (100)
	After 10 Percent Random Error	21.8% (53)	28% (126)	39.1% (121)
	After 20 Percent Random Error	22% (28)	25.2% (100)	35.9% (172)
Oregon JCP FIRE-Version (11 Items-Coefficients)	None	19.5% (77)	30.4% (107)	45.3% (116)
	After 10 Percent Random Error	21% (43)	26.1% (112)	39.2% (145)
	After 20 Percent Random Error	20.3% (21)	25.4% (88)	34.6% (191)

Results- Hypothesis #4:

- After injecting systematic (up and down) error into each risk model, the subgroup base-rates for high-risk groups will decrease and the subgroup base-rates for low risk groups will increase, showing a pattern of regression towards the mean in such rates.

Two types of systematic error are tested in the study, and each is very different. Thus, the impacts of systematic (up) error and systematic (down) error on risk tools are analyzed separately.

Systematic (up) Error and Subgroup Base-rates

The study rejects hypothesis #4. The subgroup base-rates, following the injection of systematic (up) error into the risk items, have some tendency to shift towards the mean. This is especially true for the subgroup base-rates belonging to the high-risk group. In all four-risk tools, the injection of systematic (up) error causes the high-risk groups' subgroup base-rate to decrease. For the subgroup base-rates in the low risk groups, however, this pattern is not definitive. In other words, only two of the four risk tools, Risk Device X and Oregon JCP FIRE, exhibited shifts towards the mean in the low risk groups' subgroup base-rates. For instance, systematic (up) error caused the subgroup base-rate for the low risk group to decrease to 7.1% in Risk Device X, and 13.7% in Oregon JCP FIRE (see table 4A). Thus, the study rejects hypothesis #4.

Systematic (down) Error and Subgroup Base-rates

Systematic (down) error is the only type of error that increases a risk instrument's ability to classify its high-risk individuals. So far, the injection of random and systematic (up) error into the risk tools, have been shown to compromise the systems' general capacity to separate individuals into groups with substantially different subgroup base rates. This is demonstrated by the convergence of subgroup base-rates. Conversely, systematic (down) error seems to maximize the validity of risk tools for high-risk groups, by increasing their subgroup base-rates. Instead of forcing the subgroup base rates to shrink towards the middle, systematic (down) error causes the subgroup base-rate in the high-risk group to increase. However, for the low risk groups, systematic (down) error seem to increase the subgroup base-rates, thereby reducing its ability to identify individuals that are truly low risk. Table 4A clearly shows that the subgroup base-rates increase for both the high risk groups and low risk groups when systematic (down) error is injected into the risk tools.

Table 4A: *Subgroup Base-rate Changes after Injecting Systematic Error in Four Risk Tools*

Risk Tool	Type of Systematic Error	Low (N)	Moderate (N)	High (N)
Risk Device X	None	7.2% (20)	13.9% (46)	84% (331)
	Upward	7.1% (14)	11.7% (35)	68.9% (348)
	Downward	10% (35)	27.4% (93)	85.9% (269)
Oregon JCP FIRE (3 CATS)	None	17.1% (70)	31.8% (116)	49.7% (113)
	Upward	13.7% (28)	27.6% (132)	43.5% (139)
	Downward	18.3% (82)	33.6% (123)	50.5% (94)
Oregon JCP FIRE- Version (11 Items- Burgess Method)	None	20.1% (86)	33.2% (114)	43.1% (100)
	Upward	20.5% (44)	26.4% (111)	39.5% (145)
	Downward	21.4% (101)	34.4% (117)	43.1% (82)
Oregon JCP FIRE- Version (11 Items- Coefficients)	None	19.5% (77)	30.4% (107)	45.3% (116)
	Upward	20.2% (36)	25.4% (104)	38.5% (408)
	Downward	20.6% (90)	31.4% (110)	46.5% (100)

Summary

The impact of error on misclassification in Risk Device X can be measured in two ways. First, the summation of total displacements in Table 1A and 1B shows the number of shifts that transpired after 10 percent and 20 percent error. However, the shift in cases tells a very superficial story because it doesn't directly explain how validity is affected. It only shows the number of cases that are erroneously misclassified. Second, the impact of error can be measured by comparing the failure rate for each group before and after the injection of error. This second method is preferred over the first because not only will it allow us to measure the impact of error on misclassification, but it would also allow for us to measure the impact of error on the risk instrument's overall validity. The failure rate for each group is generated by dividing the subgroup base-rate by the total number of individuals belonging to that risk group.

The impact of error is represented by both the magnitude of shift and the direction of shift. Table 2A shows that Risk Device X, Oregon JCP FIRE, Oregon JCP-Burgess, and Oregon JCP-Coefficients have a high tolerance for error. After 10 and even 20 percent of initial error into all of its items, these instruments remain valid and robust in dividing individuals into varying levels of subgroup base-rates. Though the significance tests suggest that the changes across the three categories of risk were significant at least at the .05 level, the effect sizes as evidenced by the change in subgroup base-rates suggest that the changes were inconsequential. The risk devices, thus, tolerated the injection of 10 and 20 percent error rather well. The magnitude of impact was generally mild. The subgroup base-rates for each risk group will tend to regress towards the mean after the injection of random error. That is, the recidivism rate for the low risk group will

increase, the recidivism rate for the medium risk group will stay roughly the same if it isn't already proximate to the mean, and the recidivism rate for the high risk group will decrease after the injection of random error. For instance, in Oregon JCP-Coefficients, the high risk group's failure rate experienced a minor reduction of 10.7% and the low risk group's failure rate experienced an even smaller increase of .8% after the injection of 20 percent random error. Thus, the direction of shifts resulting from the injection of error for the subgroup base rates displaces towards the mean. Also, though higher levels of error were not tested, it is generally assumed that higher levels of random error will produce greater shifts towards the mean.

The two measures used give very different estimates about the impact of error. On the one hand, small levels of error yield great levels of classification error. And on the other hand, the change in subgroup base-rates conveys that such error is inconsequential to the validity of risk tools. The new subgroup base-rates were significantly ($p<.05$) different from those of the original population after the injection of 10 percent and 20 percent random error. However, despite the enormity of the individual case displacements and the significant changes to subgroup recidivism rates, the model retains its validity in the sense that it is still able to divide individual into groups with meaningfully different base rates. The model's validity diminishes after the injection of error, as evidenced by the convergence of subgroup base rates towards the mean, and the significance of subgroup base-rate shifts, but nonetheless, the model remains a valid classification device. As such, it is learned that risk instruments generally have a high tolerance for random error.

Research Question #2: How do different distributions of cases across categories impact sensitivity?

This section examines the impact of random error on risk instruments with different distribution of cases across categories. Two hypotheses (#5 and #6) are tested to better understand the interplay between random error and the distributions of cases in categories, and their impact on the overall sensitivity of error. Thus, the first section, which concerns hypothesis #5, will look at the overall consequences of over-classifying individuals into low risk, moderate risk, and high risk. Section two compares the impact of random error on a risk instrument with three categories of risk classification and five categories of classification. This will facilitate a better understanding of how increases in risk categories impact the sensitivity of error.

Results- Hypothesis #5:

- Risk instruments that over-classify high risk individuals will be more sensitive to random error.

Risk devices may vary greatly by disproportionately assigning individuals into different risk groups, i.e. low, moderate, and high. It is thus quite important to understand how the over-classification of individuals into particular categories impact sensitivity. Three general types of distributions of cases across categories are tested: normal or the over-classification of individuals into moderate risk (see table 5A), positive or the over-classification of individuals into low risk (see table 5C), and negative or the over-classification of individuals into high risk (see table 5E). Hypothesis #5 looks at the distributions of cases across three different risk categories. For example, a normal distribution of cases in a risk instrument containing 3 risk categories will have a

significant cluster of cases in the moderate risk level, with significantly fewer cases in the low and high-risk categories (see 5A).

The study rejects hypothesis #5. Negatively skewed cases in the outcomes of the entire risk instrument do not cause the greatest increase in sensitivity. Of the three types of distributions tested, positively skewed cases produce the greatest change in net recidivism rates. After injecting 10 percent random error into the positively skewed version of Risk Device X, the recidivism rate for low, moderate, and high risk group changed 6.6%, 17.4%, and 6.6% respectively, totaling a net change of 30.6% (see Table 5D). On the other hand, the normally distributed model of Risk Device X experienced changes of 6.5%, 2.4%, and 6.6% in the low, moderate, and high-risk categories after the injection of 10 percent random error, totaling a net change of 15.5% (see Table 5B). Similarly, the injection of 10 percent random error into the negatively skewed model of risk Device X only yielded changes of 5.5%, 3.5%, and 3.4%, totaling a net change of 12.4% (see Table 5F). Thus, injecting 10 percent random error into each of the versions of Risk Device X shows that positively skewed cases in the outcomes of the entire risk device come with the highest disadvantage because it causes the highest change in subgroup base-rates.

The injection of 20 percent random error into the same three models mentioned above produces a familiar pattern, which further supports the rejection of hypothesis # 12. After imposing 20 percent error into the positively skewed model of Risk Device X, changes of 8.8%, 27.2%, and 9.4% were seen for the recidivism rates of low, moderate, and high-risk groups, with a net change totaling 45.4% (see Table 5D). The second highest change in recidivism rates was seen in normally distributed cases, with changes

of 3.5%, 1.9%, and 9.4% in low, moderate, and high risk recidivism rates, totaling 14.8% (see Table 5B). Finally, a negative distribution yields the lowest change in recidivism rates, 3.5%, 5.5%, and 5.7% respectively, totaling 14.7% (see Table 5F). Thus, positively skewed cases continue to effect the greatest change in recidivism rates.

Table 5A: *Scale Score of Risk Device X (Normal Distribution of N)*

Scale Score	N	Number of Failures	Number in Cutoff	Cutoff Base Rate (%)
7	2	0		
8	2	0		
9	3	0		
10	14	1	109 (10.9%)	5 (4.5%)
11	31	2		
12	57	2		
13	74	5		
14	93	10		
15	102	12		
16	119	12		
17	109	22	766 (76.6%)	280 (36.5%)
18	105	75		
19	90	80		
20	74	64		
21	43	37		
22	39	36		
23	23	21	125 (12.5%)	112 (89.6%)
24	13	11		
25	6	6		
26	0	0		
27	1	1		
Total	1000	397	1000	

Table 5B: *Recidivism Rates in Normal Distribution of N (Risk Device X)*

Type of Error	Low (N)	Med	High
Original	4.5% (109)	36.5% (766)	89.6% (125)
10% Random	*11%* (109)*	*34.1% (726)*	*83%* (165)*
20% Random	*7.5%* (106)*	*34.6% (722)*	*80.8%* (172)*

Table 5C: *Scale Score of Risk Device X (Positive Distribution of N)*

Scale Score	N	Number of Failures	Number in Cutoff	Cutoff Base Rate (%)
7	2	0		
8	2	0		
9	3	0		
10	14	1		
11	31	2		
12	57	2	606 (60.6%)	66 (10.8%)
13	74	5		
14	93	10		
15	102	12		
16	119	12		
17	109	22		
18	105	75		
19	90	80	269 (26.9%)	219 (81.4%)
20	74	64		
21	43	37		
22	39	36		
23	23	21	125 (12.5%)	112 (89.6%)
24	13	11		
25	6	6		
26	0	0		
27	1	1		
Total	1000	397	1000	

Table 5D: *Recidivism Rates in Positive Distribution of N (Risk Device X)*

Type of Error	Low (N)	Med	High
Original	10.8% (606)	81.4% (269)	89.6% (125)
10% Random	*17.4%* (590)*	*64%* (245)*	*83%* (165)*
20% Random	*19.3%* (548)*	*54.2%* (280)*	*80.8%* (172)*

Table 5E: *Scale Score of Risk Device X (Negative Distribution of N)*

Scale Score	N	Number of Failures	Number in Cutoff	Cutoff Base Rate (%)
7	2	0		
8	2	0		
9	3	0		
10	14	1	109 (10.9%)	5 (4.5%)
11	31	2		
12	57	2		
13	74	5		
14	93	10	269 (26.9%)	27 (10%)
15	102	12		
16	119	12		
17	109	22		
18	105	75		
19	90	80		
20	74	64		
21	43	37	622 (62.2%)	365 (58.6%)
22	39	36		
23	23	21		
24	13	11		
25	6	6		
26	0	0		
27	1	1		
Total	1000	397	1000	

Table 5F: *Recidivism Rates in Negative Skew of Cases (Risk Device X)*

Type of Error	Low (Base-rate)	Med	High
Original	4.5% (109)	10% (269)	58.6% (622)
10% Random	*11%* (109)*	*13.5% (258)*	*55.2% (633)*
20% Random	*7.5% (106)*	*15.5%* (225)*	*52.9%* (669)*

Results- Hypothesis #6:

- Increasing the number of categories to which offenders are classified will increase misclassification and the sensitivity of error.

The probability that an individual is misclassified increases when a risk assessment instrument divides individuals into more risk categories. In increasing the number of risk categories, the same range of score is partitioned into more categories of classification, thereby shrinking the range of scores that form each category of risk. As such, it is expected that marginal changes to individuals' risk scores will yield more misclassifications. To test hypothesis #6, a different version of Risk Device X, where individuals are divided into 5 categories of risk as opposed to 3, was devised. Hence, "Risk Device 5 Cat" is identical to Risk Device X in every aspect except for the number of risk categories to which individuals are assigned.

The sensitivity of error is invariably increased in risk instruments that seek to divide offenders into more category options. The study fails to reject hypothesis #6. As the number of categories to which individuals are assigned increases, the sensitivity of error also increases. Table 8A and 8B illustrate the disparity in displacements after the injection of random error into both Risk Device X with 3 categories of risk and Risk Device 5 Cat with 5 categories of risk. The aggregate displacements experienced by all 11 risk items are reported in column "Total". There are more displacements following the injecting of error experienced by Risk Device 5 Cat. While 10% random error yielded 133 upward case shifts in Risk Device X, the same error yielded 218 case shifts in Risk Device 5 Cat. And for downward case shifts, there were 111 and 174 shifts,

respectively. Likewise, 20% random error yielded 218 upward case shifts in Risk Device X and 331 upward case shifts in Risk Device 5 Cat. The same pattern is seen for downward case shifts following the injection of 20% random error, 174 and 217 respectively. Increases in the number of categories comprising a classification instrument will increase the sensitivity of error.

Table 6A: *Misclassifications (up) after Injecting Random Error into Risk Device X and Risk Device 5 Categories*

	1 to 2	2 to 3	3 to 4	4 to 5	1 to 3	1 to 4	1 to 5	2 to 4	2 to 5	3 to 5	Total
A	*56*	*64*	*N/A*	*N/A*	*13*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*133*
B	*85*	*109*	*N/A*	*N/A*	*26*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*220*
C	40	41	35	48	12	8	1	7	6	20	*218*
D	44	69	50	56	27	13	3	25	12	32	*331*

Note: Table H8A describes cases that were shifted to a higher risk level.
Row A: case shifts following 10% random error in Risk Device X.
Row B: case shifts following 20% random error in Risk Device X
Row C: case shifts following 10% random error in Risk Device 5 Cat
Row D: case shifts following 20% random error in Risk Device 5 Cat

Table 6B: *Misclassifications (Down) after Injecting Random Error into Risk Device X and Risk Device 5 Categories*

	5 to 4	4 to 3	3 to 2	2 to 1	5 to 3	5 to 2	5 to 1	4 to 2	4 to 1	3 to 1	Total
A	*N/A*	*N/A*	*57*	*50*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*4*	*111*
B	*N/A*	*N/A*	*65*	*62*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*12*	*139*
C	26	39	37	39	6	1	1	11	3	11	*174*
D	42	47	35	49	7	4	5	11	3	14	*217*

Note: Table H8B describes cases that were shifted to a lower risk level.
Row A: case shifts following 10% random error in Risk Device X.
Row B: case shifts following 20% random error in Risk Device X
Row C: case shifts following 10% random error in Risk Device 5 Cat
Row D: case shifts following 20% random error in Risk Device 5 Cat

The findings are consistent with hypothesis #6. Instruments that seek to assign individuals to more risk categories will have a lower tolerance for error. The range of scores comprising each risk category will condense as the number of risk categories increase. And as the range of scores shrink, each score point difference will yield a more robust impact. Put differently, one point of change in a broad score scale would most likely not yield any significant changes. However, when additional categories are included, the range of score for each classification also shrinks to make room for the new categories. The total range of scores does not increase or adjust to the added categories. In effect, the range of scores for each individual category will shrink to accommodate more categories of risk. For example, Table 14 (see p. 94), which shows the distribution of scores for Risk Device 5 Cat, shows how the scale score comprising each category is reduced following the inclusion of more classification categories. It is evident, thus, that the reduction in the range of scores that form a risk category is negatively correlated with the sensitivity of error. The likelihood that initial error will materialize into classification error can be minimized by reducing the number of risk categories to which individuals are assigned in a risk instrument, thereby expanding the range of scores for each risk group. So far, several separate analyses in the study have suggested that the range of scores is critical to the issue of sensitivity.

Up to this point, the discussion about hypothesis #6 is based primarily on comparisons between the collective displacements of cases before and after the injection of error. However, such discussions about case displacements are peripheral at best because it does not tap into the issue of validity. The following table will explain how the results are in support of hypothesis #6 by reviewing base-rate relationships. As was

discussed in the literature review, Baird (2009) recommends that the validity of risk instruments should be evaluated based on their relevant subgroup base rates. The individual risk groups discerned by a particular risk instrument should have demonstrated relationship to recidivism. In other words, the failure rate for each group should be commensurate of the group's designated propensity (e.g., low risk has low failure rate, high risk has higher failure rate). Proportionality is the key to obtaining validity. Thus, in this section, the discussion of hypothesis #6 will be based on the impact that the number of categories has on subgroup base-rates.

The risk instruments, whether it has three or five risk categories, seem to aptly retain their validity after the injection of error. Movement of cases from one risk category to another is inevitable, as was shown in Tables 6A and 6B. However, the validity does not seem to be reduced much after reviewing their respective subgroup base-rates. Table 6C displays the subgroup base-rates for each risk category before and after the insertion of random error into Risk Device 5 Cat. Risk Device X, which has three categories of risk classification, seems to retain its validity quite well, despite the injection of 10 and 20 percent random error (see Table 3A on p.122). The general direction of the shift of the base rate is towards the lower end. And the magnitude of impact seems to not affect the instrument's ability to effectively separate individuals into groups with varying fail rates. Subsequent the injection of 10 percent error, the base rates in some instances drops 15 percent, as in the high-risk category. However, despite this drastic shift in subgroup base rates, the instrument withstands error well and continues to effectively divide individuals into groups with meaningfully proportional subgroup base-rates.

Next, looking at Table 6C, which displays the subgroup base-rates for Risk Device 5 Cat, it is evident that the transfer of error is similarly resisted. After 10 and 20 percent of error, the subgroup base-rate shifted towards the lower end, just as it did in Risk Device X. Similarly, the cases belonging to the lower end of the spectrum will increase. In other words, the subgroup base-rates for each group will regress towards the median in a proportionate manner. Since the change is proportionate and affects all categories equally, much of the instrument's validity will be retained, even upon higher levels of injected error. The analysis does not test the impact of higher levels of error, but it is speculated that higher dosages of error will only bring the collective subgroup base-rates closer to the median.

The next logical question is: at what point does the injection of error begin to adversely affect the validity of the risk instrument. The analytic procedure used in the current study does not allow us to directly answer this question. Higher levels of error will need to be tested to understand exactly how much error is required to bring the instrument over its tipping point. The tipping point is the moment when a moderate risk group begins to experience a failure that is higher than high risk group or lower than the low risk group. However, such goal to ascertain the tipping point of any specific instrument is senseless and unnecessary because such point will vary greatly depending on a multitude of properties, such as range of scores, number of risk categories, cut-offs, subgroup base-rates, etc. What could be inferred from the study, however, is that the increase of risk categories in a risk instrument will hasten the tipping point. Thus, risk devices with more risk categories will have a positive relationship with the sensitivity of error.

Table 6C: *Subgroup Base-rates for Risk Device 5 Cat*

	Low (Base-rate)	*Low/Mod*	*Moderate*	*Mod/High*	*High*
Original	*5.4% (183)*	*11.2% (195)*	*14.9% (228)*	*86.5% (195)*	*88.4% (199)*
10% Random Error	*11.3%* (176)*	*14.1% (191)*	*25.1%* (223)*	*60%* (170)*	*80%* (240)*
20% Random Error	*10.7%* (167)*	*15.2%* (164)*	*29%* (217)*	*51.4%* (208)*	*75.4%* (244)*

Research Question #3: How do different distributions of cases across risk variables affect the sensitivity of error?

This section is an attempt to explore the different types of risk variables and their impact on sensitivity. Five hypotheses that make up this section: H7) skews, H8) kurtosis, H9) number of categories, and H10) weights are answered by focusing primarily on the findings obtained from the injection of error into Risk Device X. Finally, hypothesis #11 examines the impact of systematic error across these four types of variables.

Results- Hypothesis #7:

- Negatively skewed risk items will have the greatest impact on increasing the sensitivity of error, when compared to normally and positively skewed risk items.

The study rejects hypothesis #3, which states that negatively skewed risk items will have the greatest impact on increasing the sensitivity of error. Quite the opposite, the injection of random error into positively skewed items increases more misclassification then the injection of random error into normally and negatively skewed items. The measurement of classification error can be calculated by counting the net shifts, as shown in Table 7A. However, the measurement of impact is better accomplished by calculating the net change in sub-group base rates (see Table 7C). The reason subgroup base-rate changes are preferred over simple case shifts is because the validity of a risk device is predominantly determined by its competence in dividing groups with varying subgroup base-rates. The quantifying of case shifts is a superficial method of measurement that is irrelevant to the overall validity of the risk instrument.

In explaining the results for hypothesis #7, it is important to speak about the impact of error on each type of skew, respective to each other. As was mentioned earlier, the injection of error across risk items with varying skews roughly produces the same number of misclassified cases. For example, Item A (negatively skewed), Item B (positively skewed), and Item D (normally skewed) in Table 7A shows displacements/misclassifications of 34 cases, 34 cases, and 35 cases respectively, after the injection of random error into Risk Device X. Thus, the difference in the number of misclassifications across different skews is inconsequential. However, different skews create noticeable differences in the subgroup base-rates.

Table 7C, which displays the subgroup base-rate changes for risk items 1, 3, and 4, before and after the injection of 10 percent random error, provides support that the inclusion of positively skewed items has the greatest impact on increasing the sensitivity of error. Using row 1 in Table 7E and 7F as a baseline for comparison, which displays the actual subgroup base-rates in Risk Device X prior to the injection of error, it can be seen that Item B produces the greatest subgroup base-rate changes. But before we start calculating the net change in subgroup base rates across the columns in the following tables (7C and 7D), it is important to first understand how the measurement of sensitivity is conveyed numerically. The diminution of sensitivity is operationalized as the reduced ability to separate individuals into groups with meaningfully different subgroup base-rates. As such, reduced sensitivity is measured by the degree to which subgroup base-rates from extreme ends converge towards the middle. We can, thus, measure decreased validity or increased sensitivity of error by calculating the total subgroup base-rate changes for the high-risk group and low risk group, leaving out the middle category. The

subgroup base-rate change for the middle categories can be ignored in this calculation since it is a neutral category.

Using the method of measurement mentioned above, the findings suggest that positively skewed items are the most sensitive to error and that negatively skewed items are the least sensitive to error. After injecting 10 percent random error into each item in Risk Device X, item B, which is positively skewed, experiences the greatest degree of convergence towards the middle (see table 7C). That is, the combined subgroup base-rate changes for both the low-risk and high-risk groups are greatest, roughly 3.5 percent. As for Item A and Item D, the total percentage of convergence is .9 percent and 2.5 percent, respectively (see Table 7C). A similar pattern is seen after the injection of 20 percent random error in to Risk Device X. The net change for Items A, B, and D is .6 percent, 5.9 percent, and 2.9 percent (see Table 7D). Once again, Item B experiences the greatest level of convergence in terms of subgroup base-rate change. Note, item C is removed from the table and analysis because it belongs to a different family of risk items that which is weighed disproportionately more.

The differential impact caused by different skews on base-rates is analyzed. The logic behind the construction of a valid risk device is premised on creating risk classification categories with subgroup base-rates that would progressively increase with the level of risk. Naturally, the majority of failures would cluster towards the high-risk end. Given this general distribution of failures, displacements towards the lower-risk end will have the biggest impact on altering the subgroup base-rates in a way that would detrimentally affect the validity of risk devices. Following this same line of logic, negatively skewed items, which create clusters on the right hand side or high-risk group,

pushes cases towards the lower end. In essence, the skew of the risk items determines both the general shift of cases and the general shift in subgroup base-rates.

The inclusion of positively skewed and normally distributed risk items reduces the sensitivity of error. For mostly the same reason, positively skewed risk items will do the opposite and reduce sensitivity because it pushes the majority of cases and base-rates toward the high-risk end, where failures rightfully belong. Though there is a tipping point to which the exodus of cases to the high-risk end will yield unsalvageable changes, generally, risk instruments are much more tolerable of displacements towards the high-risk end. As for normally distributed risk items, the shifts are multi-directional, thereby creating somewhat of a canceling out effect within risk items. For these reasons, it is evident that negatively skewed items are more potent than its positive and normal counterparts in increasing sensitivity.

Table 7A: *Number of Misclassifications After Injecting 10% Random Error*

Risk Tool	Low to Moderate	Moderate to High	Low to High	High to Moderate	Moderate to Low	High to Low	Total
Item 1	6	3	0	14	11	0	34
Item 2	14	16	0	2	2	0	34
Item 3	9	16	7	14	16	4	56
Item 4	6	12	0	6	11	0	35
Item 5	4	4	0	5	5	0	18
Item 6	11	7	0	2	3	0	24
Item 7	4	2	0	12	5	0	23
Item 8	0	1	0	10	10	0	21
Item 9	5	6	0	12	5	0	28
Item 10	11	19	2	6	7	0	45
Item 11	15	15	0	1	3	0	34
Items Total	56	64	13	57	50	4	244

Note: Table 7A illustrates the direction and magnitude of shifts after 10 percent random error was injected, while Table 7B illustrates the same properties of the same dataset after the injection of 20 percent error. The items on the far left column represent the number of risk items contained in the risk instrument. Both tables (7A and 7B) are describing the same risk instrument and set of data. The primary difference between the two tables is their disparity in terms of the level of error that was injected.

Table 7B: *Number of Misclassifications After Injecting 20% Random Error in Risk Device X*

Items to which error is injected	Low to Moderate	Moderate to High	Low to High	High to Moderate	Moderate to Low	High to Low	Total
Item 1	11	3	0	32	23	0	69
Item 2	24	31	0	6	4	0	65
Item 3	12	17	10	17	8	2	66
Item 4	9	22	0	12	21	0	64
Item 5	6	7	0	13	9	0	35
Item 6	25	17	0	6	4	0	52
Item 7	4	3	0	21	11	0	39
Item 8	4	3	0	22	26	0	55
Item 9	13	16	0	20	8	0	57
Item 10	19	45	5	11	10	0	90
Item 11	23	29	0	4	4	0	60
Items Total	85	109	26	65	62	12	359

Note: Table 7A illustrates the direction and magnitude of shifts after 10 percent random error was injected, while Table 7B illustrates the same properties of the same dataset after the injection of 20 percent error. The items on the far left column represent the number of risk items contained in the risk instrument. Both tables (7A and 7B) are describing the same risk instrument and set of data. The primary difference between the two tables is their disparity in terms of the level of error that was injected.

Table 7C: *Subgroup Base-rate Change in Items 1,2,4 after Injecting 10% Random Error in Risk Device X*

Item to which Error is injected	Low (Base-rate)	Moderate	High	Total Change (Low + High)
Original	7.2% (20)	13.9% (46)	84% (331)	N/A
Item 1	.6%	1.5%	.3%	.9%
Item 2	.3%	.3%	3.2%*	3.5%
Item 4	.2%	1.4%	2.3%	2.5%

Note: Row "Original" provides the baseline for comparison. It shows the subgroup base-rates prior to any disruption by error. *The asterisks indicate significance of .05 or smaller in the change in subgroup base-rate. The percentages are obtained by finding the difference between the new subgroup base-rate (after injecting error into specified risk item) and the original subgroup base-rate. For the purpose of this analysis, Item 1 is negatively skewed, Item 2 is positively skewed, and Item 3 is normally distributed.

Table 7D: *Subgroup Base-rate Change in Items 1,2,4 after Injecting 20% Random Error*

Item to which Error is injected	Low (Base-rate)	Moderate	High	Total Change (Low + High)
Original	7.2% (20)	13.9% (46)	84% (331)	N/A
Item 1	0%	5.4%*	.6%	.6%
Item 2	.6%	.5%	5.3%*	5.9%
Item 4	0%	1.6%	2.9%	2.9%

Note: Row "Original" provides the baseline for comparison. It shows the subgroup base-rates prior to any disruption by error. *The asterisks indicate significance of .05 or smaller in the change in subgroup base-rate. The percentages are obtained by finding the difference between the new subgroup base-rate (after injecting error into specified risk item) and the original subgroup base-rate. For the purpose of this analysis, Item 1 is negatively skewed, Item 2 is positively skewed, and Item 3 is normally distributed.

Table 7E: *Subgroup Base-rates by Risk Category and Risk Item after injecting 10% Random Error Risk Device X*

Item to which Error is injected	Low Base-rate% (N)	Moderate	High
Original	7.2% (276)	13.9% (330)	84% (394)
Item 1	7.8% (281)	15.4% (336)	84.3% (383)
Item 2	7.5% (264)	14.2% (330)	80.8%* (408)
Item 3	9.2% (280)	15.8% (321)	80.2%* (399)
Item 4	7.4% (281)	15.3% (319)	81.75% (400)
Item 5	7.5% (277)	15.1% (330)	82.9% (393)
Item 6	6.7% (268)	14.4% (333)	82.9% (399)
Item 7	7.5% (277)	15.6% (339)	84.1% (384)
Item 8	6.9% (286)	16.1% (329)	84.1% (385)
Item 9	6.5% (276)	16.3% (336)	83.5% (388)
Item 10	8.8% (270)	12.4% (321)	81.4% (409)
Item 11	6.8% (264)	14.3% (328)	81.3% (408)
Items Total	10.8%* (276)	22.1%* (329)	71.7%* (410)

Note: Row "Original" provides the baseline for comparison. It shows the subgroup base-rates prior to any disruption by error *The asterisks indicate significance of .05 or smaller in the change in subgroup base-rate. The numerator represents the new subgroup base-rate. The denominator represents the number of individuals that fall into each category. The percentages are obtained by finding the quotient.

Table 7F: *Subgroup Base-rates by Risk Category and Risk Item after Injecting 20% Random Error in Risk Device X*

Item to which Error is injected	Low Base-rate% (N)	Moderate	High
Original	7.2%	13.9%	84%
	(276)	(330)	(394)
Item 1	7.2%	19.3%*	84.6%
	(288)	(347)	(365)
Item 2	7.8%	14.4%	78.7%*
	(256)	(325)	(419)
Item 3	7.1%	17%	79.8%*
	(264)	(334)	(402)
Item 4	7.2%	15.5%	81.1%
	(288)	(308)	(404)
Item 5	7.5%	16.2%	82.9%
	(279)	(333)	(388)
Item 6	6.2%	14.7%	81.7%
	(255)	(344)	(405)
Item 7	8.1%	17.5%	83.5%
	(283)	(341)	(376)
Item 8	6.3%	18.9%*	84.2%
	(298)	(327)	(375)
Item 9	6.2%	17.9%*	81.7%
	(271)	(339)	(390)
Item 10	8.7%	14%	76.4%*
	(262)	(305)	(433)
Item 11	5.9%	15.4%	78.7%*
	(287)	(324)	(419)
Item All	11.2%*	25.5%*	64.3%*
	(239)	(309)	(452)

Note: Row "Original" provides the baseline for comparison. It shows the subgroup base-rates prior to any disruption by error *The asterisks indicate significance of .05 or smaller in the change in subgroup base-rate. The numerator represents the new subgroup base-rate. The denominator represents the number of individuals that fall into each category. The percentages are obtained by finding the quotient.

Results- Hypothesis #8:

- Increasing the pointiness of distributions in risk items will increase the amount of error or misclassifications towards a particular direction.

Asides from the different types of skews (negative, positive, normal) that impact sensitivity, variables may differ on their pointiness. Pointiness refers to the height of the skews, which could either be leptokurtic or platykurtic. It is speculated that the peakness of a distribution has a major impact on the magnitude of shifts or the number of cases that will be displaced. The study fails to reject Hypothesis #8, which posits that increases in any skew's kurtosis will create increases in displacements towards a direction.

To test hypothesis #8, the number of misclassifications that is produced after the injection of error is compared between pointy and flat risk items. Five pointiest and flattest risk items that are also positively skewed are selected from Oregon JCP FIRE for comparison, and their displacements are averaged. Table 8 (see p.85) shows the distributions of risk items that together comprise the Oregon Risk Instrument. Items 1, 3, 7, 22, and 26 are selected to represent those that are platykurtic, while items 4, 14, 15, 19, and 20 are selected to represent those that are most peaked. In Table 8A, the number of misclassifications for the selected risk items is shown. The study fails to reject hypothesis #8. Increases in the kurtosis of a skew do not produce more misclassifications, but rather, it dictates the magnitude of the direction of the displacements. In other words, 10 percent random error will create similar quantities of misclassifications across all risk items. The direction of the misclassifications is determined by the direction of a skew, and the degree of change towards a direction is determined by the kurtosis.

As illustrated in Table 8A, five pointy risk items selected exhibit a higher level of disproportionality. The higher the disproportionality of the ratios, the greater the direction of misclassification is shown. For example, pointy risk item #4 has a ratio of 47:5. This means that 47 of the misclassified cases shifted to a higher risk category and only 5 cases shifted to a lower risk category. Conversely, risk item #1, which has a flatter distribution, exhibits a much lower ratio of 25:16 for cases that shifted to a higher risk designation and for cases that shifted to a lower risk designation. The same patterns are seen for the other selected risk items, giving support to hypothesis #4. Large discrepancies between these two numbers (case shift up and case shift down) indicate that the magnitude of the shifts is also very large. Thus, the injection of random error into pointy risk items will produce greater levels of shifts towards a specific direction, while the injection of random error into flatter risk items will create milder shifts.

The injection of random error into such items lends support to the hypothesis that a high kurtosis in the distributions of the cases for individual risk items will increase the magnitude of displacements. The total displacements experienced by both pointy and flat risk items are roughly similar. The main difference is in how the kurtosis impacts the direction of shift.

Table 8A: *Number of Misclassifications (for select items) after 10% Random Error into Oregon JCP FIRE*

Items to which Error is Injected	Low to Low Moderate	Low Moderate to Moderate	Moderate to High	High to Moderate	Moderate to Low Moderate	Low Moderate to Low	Shift Up (total)	Shift Down (total)
Item 1	7	12	6	11	4	1	25	16
Item 3	16	11	8	8	4	1	35	13
Item 7	7	6	10	7	2	3	23	12
Item 22	9	20	6	3	6	1	35	10
Item 26	12	0	0	5	3	5	12	13
Item 4	14	14	19	1	4	0	47	5
Item 14	17	18	14	2	0	1	49	3
Item 15	15	14	14	1	3	3	43	7
Item 19	13	20	11	0	2	2	44	4
Item 20	12	13	11	2	3	0	36	5

Note: Items 1, 3, 7, 22, and 26 are selected to represent those that are platykurtic. Items 4, 14, 15, 19, and 20 are selected to represent those that are most peaked.

Results- Hypothesis #9:

- Dichotomous risk items will have a greater impact on the sensitivity of error that will risk items with more category options.

The inclusion of dichotomous risk items moderates the sensitivity of error, thereby the study fails to reject hypothesis #7. In fact, an opposite trend occurs when dichotomous variables are injected with error. Fewer displacements are seen from dichotomies after the injection of random error, compared to 3-leveled and 4-leveled risk items. Rows 5, 6, and 7 from Table 7A (see p.144) display the direction and magnitude of the case shifts following the injection of 10 percent random error. Item 5 is normally distributed. Item 6 is negatively distributed. And finally, Item 7 is positively distributed. The net shifts in Items 5, 6, and 7 following the injection of 10 percent random error are 18, 24, and 23, respectively (see table 7A). These numbers, which represent the total misclassified cases, are much smaller than those displayed for 4-leveled risk items (1-4) and 3-leveled risk items (8-11). Similar patterns are seen in Table 7B (see p. 145), which displays the displacements of cases following the injection of 20% random error. Likewise, Items 5, 6, and 7 experienced minimal levels of displacements, 35, 52, and 39 respectively. These numbers are much smaller than those of any other risk item, and suggest that the transfer of error is significantly reduced by the inclusion of dichotomous risk items.

Since dichotomous risk items yield the fewest displacements, it can be logically assumed that risk items that are made up of more categories will yield the greatest number of displacements. Increases in the number of risk options available in a risk item

will increase the likelihood and sensitivity of error. The findings as displayed in Tables 7A and 7B (p. 144-145) are suggestive of a positive relationship between number of risk options and level of sensitivity. Sensitivity could be measured by the magnitude of displacements following the injection of error. Clearly, four leveled risk options generate the greatest number of case shifts when compared with those of dichotomous and 3-leveled risk items. Four-leveled risk items 1, 2, 3, and 4 consistently display higher numbers than their counterparts, 34, 34, 64 and 35 relatively. The findings are generally consistent with the rejection of hypothesis #9, which states that dichotomous risk items or items with fewer risk options will have a greater impact on sensitivity than will risk items with more category options.

The subgroup base-rates, however, do not offer conclusive evidence, neither supporting nor rejecting hypothesis #9. Risk items 5, 6, and 7 in Tables 7E (see p. 147) and 7F (see p.148) display the subgroup base-rate change within the dichotomous risk items. From these tables, there isn't sufficient indication pointing to either support or rejection of the hypothesis #7, which states dichotomous risk items produce more error than 4-leveled and 3-leveled risk items. This is largely because no visible pattern among the subgroup base rates emerged as being definitive. If dichotomous risk variables consistently yielded more or less change in subgroup base-rates, inferences could be made about its impact on sensitivity. But instead, the base-rates oscillate and adhere to no particular pattern, rendering the interpretations based solely on Tables 7E and 7F (p. 147-148) inadequate.

The lack of significant patterns and findings in the subgroup base-rates, however, is not completely uninformative. Several meaningful inferences could be drawn. First, it

is plausible that the impact of risk options is relatively small when compared with other factors, causing no visibly significant patterns in the subgroup base-rates. In other words, comparing shifts in base rates may not be a robust enough test to detect small changes. This makes much sense because subgroup base rates are much more tolerable of error than are individual case shifts. Thus, interpreting subgroup base rates from Tables 7E and 7F is not a suitable method to measure sensitivity because the impact of risk options is too small to be detected using such method.

Second, the risk instrument might be constructed in such as way that specific variables could not be isolated or controlled. There exist both aggravating (increases sensitivity) and mitigating (decrease sensitivity) factors that, ideally, should be measured individually. However, Risk Device X is constructed in a way that does not allow for the individual manipulation of variables. Thus, it is possible that dichotomous risk items, which are considered mitigating factors, are inextricably intertwined with aggravating factors such as negative skews. In such a scenario, the more dominant of the two opposing factors will likely overshadow the impact produced by the less dominant factor, thereby lending itself to the appearance that effects were small or nonexistent at all. Thus, there may be cancel-out effects that the analysis fails to identify due to its inability to effectively separate different variables, allowing variables to be confounded.

Third, the effects of dichotomous risk items may be so small that it becomes overshadowed by the discrepancy that which random draws create. The process of selecting cases to which error is injected relies heavily on random draws to insure objectivity. However, it is possible that the difference caused by such random draws will dwarf the effect caused by dichotomous risk items. Thus, the effects of dichotomous risk

items may be weak or the difference caused by random draws is too big. More likely than not, the effects of dichotomous risk items do not produce strong enough effects to surpass the effects of random distortions. In conclusion, the lack of visibly significant impact on the subgroup base-rates could be more insightful and meaningful than it appears.

Dichotomous risk variables are least sensitive to random error due to their simplicity and the mildness of alternatives. Since the possible outcomes for dichotomies in Risk Device X are coded either 0 or 1, the severity of an actual mistake or error is minimal. Error in dichotomies, thus, only alter total risk scores by increments of one, either 1 point up or 1 point down. On the other hand, the impact of mistakes and error in items with four categories could potentially alter individual risk scores in a drastic way. For instance, item 4 (see table 7A, p. 144) contains four categories, and they are attached to scores of 0, 1, 2, and 3. In situations where random error would change an individual's placement from 0 to 3, the impact of the error would have caused a four-point difference. For this reason, dichotomies can better tolerate random error. The severity of potential mistakes is much smaller in these situations.

Results- Hypothesis #10:

- Increasing the weight of individual risk items will increase the sensitivity of error.

Risk items can vary greatly in terms of the score points they individually contribute to the overall risk function. Though some instruments, for the purpose of increased manageability, are strictly designed so that every risk item contributes the same score to the entire function (as in most instruments scored base on the Burgess Method),

some other instruments may contain an array of risk items that would differentially contribute to the risk function. It is hypothesized that the injection of error into such risk items, whose score-point contribution is greater than the score-point contribution of its counterparts, will have a greater impact on misclassification by increasing the sensitivity of error.

Items 3 and 4 from Risk Device X are specifically designed to test this hypothesis; the weight of the risk items and their options is much higher than that of other risk items. The Categories column from the Table 5 (see p. 81) displays the weight distributions across all 11 risk items. To test hypothesis #5, we need to pay particular attention to the relative weight distributional schematic of items 3 and 10, which is much higher than that of other risk items. While the generic schematic of the risk items follow the "0,1,2,3" pattern and categories moving from low to high increases by increments of 1, Item 3 and 10 follow the schematic of "2,4,6,8" with increments of 2. This intended disparity in their weight distributions will tell us the weights of items impact sensitivity.

The results show that the point score comprising each risk item will greatly impact the sensitivity of error. Table 7A (see p. 149) and 7B (see p.150) show that Item 3 and Item 10 continuously produce more displacements than all other risk items, after the injection of 10% and 20% random error into all items evenly. Ten percent random error into Items 3 and 10 produced 56 and 45 case displacements, while twenty percent random error into the same risk items produced 60 and 90 case displacements, respectively.

Tables 7A and 7B further suggest that raising the point score of a single risk item will have an impact on sensitivity that far surpasses those of other risk properties. For example, the type of skew, peakness of the skew, or the number of risk options in the items all affects the sensitivity of error. But, the displacement of cases from these factors is much smaller than the displacement of cases caused by increasing the point score of a risk item. The number of cases displaced in items 3 and 10 far exceeds those of any other risk item. This strongly indicates that the number of points making up individual risk items has the greatest impact on sensitivity.

The impact of score points on the subgroup base-rates tells a similar story. In particular, Items 3 and 10 from Tables 7E (see p. 147) and 7F (see p.148) show that increases in the weight of risk items have the biggest impact on sensitivity, followed by deep positive skews. To help examine the relationship of the weight of risk items and sensitivity, the weight of two risk items are increased. Risk items 3 and 10 contain options that increases by two points while all other risk items go up in increments of one. After the injection of 10 percent and 20 percent error, both of these risk items produced more changes in base-rates than all other risk items.

The changes are better realized after some further explanation of the instrument and the analytical procedure. For all the risk items, the general tendency of shifts in subgroup base-rate, after the injection of error, is one that regresses towards the mean. For instance, the recidivism rate for the low-risk group will increase, the middle group will remain unchanged, and the high-risk group will experience decreases in recidivism rates. The best way to compare risk items and their impact, thus, is to examine the drops in high-risk recidivism rate, increases in low-risk recidivism rate, and change in mod-risk

recidivism rate. Measuring the change in each risk category and summing them up will give a clear indication in the total magnitude of change. Next, comparisons should only be made between items falling under the same category. Again, looking at the Risk Device X Description Table 5 (see p. 81), items 1 to 4 are similar in terms of risk categories and should be grouped together. The same grouping should be applied to items 5 to 7 and also items 8 to 11. Comparing items to similar items have two major benefits. First, it provides some baseline for comparison. Second, it will naturally control for variables, giving us more confidence that disparities are caused by known differences.

Now, the Tables 7E and 7F (p.147-148) can be interpreted. Subtracting from the original base-rates in Row 1, the magnitude of change can be calculated. Note, Row 1 with the original base-rates is important because it lets us know the respective base rates prior to the injection of error. After subtracting subgroup base-rates in Item 3 from the original base-rates, we can see that there was a 2 percent increase in the failure rate for the low-risk group, 1.9 percent increase in the failure rate for the mod-risk group, and a 4 percent decrease in the failure rate for the high-risk group, totaling 7.9 percent change in the recidivism rate after injecting 10 percent error. We can see that the total change in recidivism rate for similar items 1, 2, and 4 is much smaller, 2.4 percent, 4.6 percent, and 3.9 percent respectively.

Next, changes in the recidivism rate for item 10, with similar weight distributions as item 3, should be analyzed to confirm findings. Again, risk items should only be compared to similar risk items, and in this case, Item 10 should only be compared to Items 8, 9, and 11. Item 10 experienced 1.6 percent change in the recidivism rate in the

low-risk group, 1.5 percent change in the recidivism rate in the mod-risk group, and 2.6 percent change in the recidivism rate in the high-risk group, totaling 5.7 percent change after the injection of 10 percent error. The total changes experienced by Risk Items 8, 9, and 11 are 2.8 percent, 3.8 percent and 1.5 percent respectively. This is another clear indication that the weight of the risk items strongly impacts the sensitivity of error in risk instruments.

The same general pattern is seen after the injection of 20 percent error. From Table 3A (p. 122), we can see that the total change in subgroup base-rates for Item 3 was 8.4 percent, compared to 6 percent, 5.6 percent, and 4.5 percent change in Items 1, 2, and 4. The comparison between the total subgroup base-rate change of Item 3 and that of its counterparts, Items 1, 2, and 4, indicates that doubling the increment of points in the options will increase the sensitivity of error. Similarly, for Item 10, the total change in subgroup base-rates was 9.2 percent, compared to 6.1 percent, 8.3 percent, and 8.1 percent in Items 8, 9, and 11. Item 10 is another risk item that contains increments of two points in its options. After the injection of 20 percent error, Item 10 produced more changes in base-rates than did other similar risk items. Thus, the study fails to reject hypothesis # 10. The weight of risk items has a positive relationship with sensitivity of error in the instrument.

Results- Hypothesis #11:

- Random error will have a smaller impact on misclassification than would systematic error.

There are generally two types of error, systematic and random. Though the study focuses on both types, it is assumed that most error within risk classification instruments is random. Unfortunately, the current state of knowledge is lacking in this area of research. The existence of random error and its prevalence are speculative at best. The influence of systematic error should not be completely dismissed, however. The current study assumes that both random error and systematic error are equally important. Thus, despite their prevalence or lack there of in the real world, equal attention is paid to each in the current study.

Hypothesis #11, which states that random error will have a smaller impact on misclassification than would systematic error, is rejected. Though the impact of systematic and random error can be mitigated or aggravated by contextual factors, systematic error, in general, creates less classification error. However, without knowing the specific circumstances or specific properties contained in a risk device to which systematic and random are injected, the relative impact of each error type cannot be determined definitively.

Adding random error into individual risk items with varying distributions yields consistently similar results. For example, when comparing the number of shifted cases in Risk Device X following the injection of random and systematic error, it is evident that random error produces more misclassifications in all eleven risk items, both individually and collectively. Table 11A, which shows the impact of systematically upward error in Risk Device X, and Table 11B (p. 145), which shows the impact of systematically downward error in the same Risk Device X, show that fewer case shifts are consistently recorded when compared with the impact of random error (see Table 7A, p.144).

Furthermore, the total displacements collectively experienced by all risk items are higher following the injection of random error. Thus, random error has a greater impact on classification error across the eleven risk items despite their differences in with distributions and risk options. Regardless of the type of distribution (positive, negative, or normal) or risk options contained in each risk item (dichotomous, 3 and 4 options), systematic error, whether up or down, causes smaller classification error.

Systematic error causes lower levels of misclassifications because it affects smaller portions of cases. When random error is injected into Item E (see Risk Device X), for example, cases from lower options could move up and cases from upper options could move down. Thus, all of the selected cases are affect by random error. However, systematic error could only affect a portion of the cases. If systematic (up) error is injected, cases in upper levels remain the same because there is no place to which these cases could displace. Thus, only cases in the lower options are affected by such error. In essence, systematic error, whether it is up or down, effects change only in a portion of the cases. Random error, on the other hand, affects all of the selected cases, causing more changes to individual risk scores, thereby increasing the number of misclassified cases.

Table 11A: *Number of Misclassifications After Injecting 10% Systematic Error (up) in Risk Device X*

Items Error Injected In	Low to Moderate	Moderate to High	Low to High	High to Moderate	Moderate to Low	High to Low	Total
Item 1	4	5	0	0	0	0	9
Item 2	13	8	0	0	0	0	21
Item 3	18	16	0	0	0	0	34
Item 4	12	9	0	0	0	0	21
Item 5	3	8	0	0	0	0	11
Item 6	6	7	0	0	0	0	13
Item 7	1	2	0	0	0	0	3
Item 8	4	1	0	0	0	0	5
Item 9	8	7	0	0	0	0	15
Item 10	12	22	0	0	0	0	26
Item 11	10	9	0	0	0	0	19
Items Total	73	104	7	0	0	0	184

Note: Table 2A displays case shifts after 10 percent of systematically upward error is injected. No cases were recorded for downward displacement due to the nature of the impact of this type of error.

Table 11B: *Number of Misclassifications after Injecting 10% Systematic Error (down) into Risk Device X*

Items Error Injected In	Low to Moderate	Moderate to High	Low to High	High to Moderate	Moderate to Low	High to Low	Total
Item 1	0	0	0	14	5	0	19
Item 2	0	0	0	6	2	0	8
Item 3	0	0	0	15	16	0	31
Item 4	0	0	0	10	7	0	17
Item 5	0	0	0	1	8	0	9
Item 6	0	0	0	2	0	0	2
Item 7	0	0	0	7	7	0	14
Item 8	0	0	0	10	8	0	18
Item 9	0	0	0	11	6	0	17
Item 10	0	0	0	8	6	0	14
Item 11	0	0	0	0	2	0	2
Items Total	0	0	0	79	70	2	151

Note: Table 2B displays case shifts for Risk Device X after 10 percent of systematically downward error is injected. No cases were recorded for upward displacement due to the nature of the impact of this type of error.

Research Question #4: Does the expansion of scale scores in risk instruments increase the sensitivity of error?

The scale score is the range of scores from which risk categories are formed. This section seeks to understand the relationship between the scale score range and the sensitivity of error. This analysis focuses on the (H12) number of risk variables in risk tools and its relationship to the vastness of the range of score. This research question can be addressed by better understanding the factors that play a pivotal role in determining the vastness of scale scores.

Results- Hypothesis #12:

- Adding more risk items to a risk tool will decrease the sensitivity of error.

Increasing the number of risk items in risk devices will reduce the tendency for misclassification error, thereby reducing sensitivity. The underlying rationale in support of this assumption comes from the natural expansion of scores that would occur due to the addition of risk items. The expansion of scores will increase the vastness of the range of scores, thereby decreasing the risk instrument's proclivity for misclassification.

The findings fail to reject hypothesis #6. To explain the differential impact of error on risk tools with different numbers of risk items, the attention could be turned towards Oregon JCP FIRE and Oregon JCP-Burgess. In this situation, Oregon JCP FIRE represents a risk tool with more items (30 risk items), and Oregon JCP-Burgess represents a risk tool with fewer items (11 risk items). Since these two risk tools are nearly identical in every other aspect, their differences in misclassification can easily be attributed to their disparity in the number of risk items contained in each risk tool. Tables

1A and 1B (p.115) show such disparities. In fact, when random error is injected into Oregon JCP FIRE, it causes 26.3% misclassification (after 10% error) and 48.3% misclassification (after 20% error). For Oregon JCP-Burgess, the level of misclassification was 33.7% and 52.7% respectively. This means that Oregon JCP FIRE (30 items) significantly reduces the transfer of error, by 7.4% and 5.4% respectively.

The number of scores for each category is determined by the cut-offs. Increases in a category's range of scores reduce the potential misplacement of individuals caused by error because there would be more room for error. This means that each score point difference would have a lower likelihood of causing classification error.

Before discussing the study's findings on the relationship between "range of scores" and sensitivity, it is important to explain the different circumstances that directly influence the range of scores. Several conditions are directly linked to the broadness of the range of scores in a risk instrument. First, increases in the number of risk items will typically cause an increase in the range of scores. Take, for example, the Oregon Risk instrument that contains 31 dichotomous risk items, with each individual item varying from a score of 0 and 1. Here, the range of score for the entire Oregon Risk instrument is 0-31, with 31 being the highest possible total. Thus, there is a clear connection between the number of risk items and the range of score. Second, the inclusion of risk items with multiple options typically increases the range of score for a risk instrument. For instance, if a risk item divides individuals into five categories, and each category is coded in increments of 1 (ex. 0, 1, 2, 3, 4), that particular risk item could contribute up to 4 points to the total risk function score, thereby substantially expanding the score.

To test hypothesis #12, error is injected into both Oregon JCP FIRE and Oregon JCP-Burgess. Both instruments are identical and differ on one relevant component, their scale score. While the Oregon JCP FIRE contains a range of 0 to 28 (see table 7, p. 85), the Oregon JCP-Burgess contains a range of score of 0 to 11 (see table 9, p. 90). This disparity in the two instruments will give us confidence that their unique range of score directly causes the differences in changes.

The study fails to reject hypothesis #12. Increasing the range of score for a risk instrument reduces its sensitivity to error. Of all the different risk device properties that have been tested to assay their impact on sensitivity, range of score seems to be most robustly linked to sensitivity. Table 1A and 1B (see p. 122) compares the number of misclassified cases in Oregon JCP FIRE and Oregon JCP-Burgess. Whether it is injecting 10% or 20% random and systematic error, Oregon JCP-Burgess consistently produces greater levels of misclassifications. For instance, after 10 and 20 percent random error, Oregon JCP-Burgess experiences 263 and 483 total case shifts while Oregon JCP FIRE experiences 337 and 527 shifts. The differences as measured by the number of individual shifts draw a very vivid picture about Oregon JCP FIRE's ability to resist classification error.

The same disparities are found after systematic error is injected into these two models. Whether it is upward systematic error or downward systematic error, greater levels of displacements are seen in the Oregon Simulation model. For instance, injecting 10 percent upward systematic error yields 298 total misclassification in the Oregon JCP FIRE, and 332 total misclassifications in Oregon JCP-Burgess (see Table 2A, p. 117). Similarly, 10 percent downward systematic error yields 80 total misclassifications in

Oregon JCP FIRE, and 86 total misclassifications in the Oregon JCP-Burgess (see Table 2B, p. 118). In every scenario described above, the instrument with the narrower range of score (Oregon Risk Instrument) consistently produces less case shifts following the injection of error.

Summary

The current chapter focused on the statistical analysis to address research questions with regard to the sensitivity of error in offender risk classification based on risk device characteristics such as distribution of data, risk categories, and range of score. Thus, twelve hypotheses were tested by measuring case displacements, change in subgroup base-rates, and significance. Each statistical method of analysis is adept in answering different specific questions. At times, one specific type of method would be superior, while others analyses would be conducted as supplemental analyses. However, the order of importance would quickly change from one research hypothesis to the next. The analyses can be divided into three domains: 1) overall impact of error on validity, both random and systematic; 2) the impact of distribution of cases in categories on sensitivity; and 3) the impact of distribution of cases in variables on sensitivity.

H1 tested the impact of random error on validity across four risk tools: Risk Device X, Oregon JCP FIRE, Oregon JCP-Burgess, and Oregon JCP-Coefficients. It showed that small level of errors typically contribute to high levels of misclassifications. *H2* tested the impact of systematic (up and down) error on risk models. Systematic (down) error caused the least number of misclassifications, which was consistent across all four risk tools. *H3* tested the impact of random error on the validity of risk models by examining changes to subgroup base-rates. The risk instruments retained its ability to

form groups of individuals whose subgroup base-rates were distinctively and meaningfully different. Next, (*H4)* looking at the impact of systematic error on validity by examining subgroup base-rates, the validity of the risk tools did not decrease much. Thus, risk tools are said to be highly tolerant of random and systematic error.

The second section examined the impact of distributions of cases in risk categories on sensitivity. *H5* tested the different skews in the risk categories. Specifically, it tested whether instruments that over-classify high risk individuals were more sensitivity to error. The study found that such instruments would reduce the sensitivity of error. Finally, looking at the number of categories to which individuals are assigned, *H6* tested the impact of instruments with more risk categories on sensitivity. The study found that more misclassifications resulted, thereby increasing the sensitivity of error in risk tools with more risk categories.

Finally, some risk device properties that pertain to the distribution of cases in risk variables were found to have an ameliorative effect on sensitivity. The properties that were found to abate the transfer of error are: dichotomies (*H9)*; items with relatively lower weights (*H10)*; and instruments with many risk items (*H11); and* increased range of scores (*H12)*. The study found that changes to the skews (*H7)* and kurtosis (*H8)* neither increases nor decreases the tendency for misclassifications. However, changes to the skews and kurtosis does impact the subgroup base-rates, thereby affecting the validity of risk tools. In sum, risk properties can directly influence the level of sensitivity of error in risk classification instruments.

The following chapter discusses the significant research findings, interpretation of such findings, theoretical and practical implications, the limitations of the study, and recommendations for further research.

Hypotheses Summary

Research Question #1: What is the impact of error on risk assessment instruments?

Hypothesis #1: After injecting random error into entire risk models, the level of classification error is equal to the level of error that is injected into the risk items.

Rejects- Table 1A and 1B

Hypothesis #2: After injecting systematic error (up and down) into entire risk models, the level of classification error is equal to the level of error that is injected into the risk items.

Rejects- Table 2A and 2B

Hypothesis #3: After injecting random error into each risk model, the subgroup base-rates for high risk groups will decrease and the subgroup base-rates for low risk groups will increase, showing a pattern of regression towards the mean in such rates.

Fails to Reject- Table 3A

Hypothesis #4: After injecting systematic (up and down) error into each risk model, the subgroup base-rates for high risk groups will decrease and the subgroup base-rates for low risk groups will increase, showing a pattern of regression towards the mean in such rates.

Fails to Reject- Table 4A

Research Question #2: How do different distributions of cases across categories impact sensitivity?

Hypothesis #5: Risk instruments that over-classify high risk individuals will be more sensitive to random error.

Rejects- Table 5B, 5D, 5F

Hypothesis #6: Increasing the number of categories to which offenders are classified will increase misclassification and the sensitivity of error.

Fails to Reject- Table 6A, 6B, 6C

Research Question #3: How do different distributions of cases across risk variables affect the sensitivity of error?

Hypothesis #7: Negatively skewed risk items will have the greatest impact on increasing the sensitivity of error, when compared to normally and positively skewed risk items.

Rejects- Table 7A, 7B, 7C, 7D, 7E

Hypothesis #8: Increasing the peakness of distributions in risk items will increase the amount of error or misclassifications towards a particular direction.

Fails to Reject- Table 8A

Hypothesis #9: Dichotomous risk items will have a greater impact on the sensitivity of error than will risk items with more category options.

Rejects- Table 7A, 7b, 7E, 7F

Hypothesis #10: Increasing the weight of individual risk items will increase the sensitivity of error.

Fails to Reject- Table 7A and 7B

Hypothesis #11: Random error will produce less misclassification than will systematic error.

Rejects- Table 11A and 11B

Research Question #4: Does the expansion of scale scores in risk instruments increase the sensitivity of error?

Hypothesis #12: Adding more risk items to a risk tool will decrease the sensitivity of error.

Fails to Reject- Table 1A and 1B

Chapter 6

Discussion and Conclusions

The primary purpose of this dissertation was to explore the relationship between initial error in risk assessment information and final classification error through the use of an experimental design. Since research on the topic of sensitivity of error in offender risk classification instruments is virtually non-existent, there is no ongoing framework from which the current study could follow. Being the first to break ground on this type of research was greatly auspicious as it is debilitating. The pioneering nature of the study makes the study and its findings unique, interesting, and momentous. However, the lack of an existing structure, on which the study could piggyback, renders the study exploratory.

To address this issue, the study started with a very small set of general questions. From there, more detailed questions and hypothesis were generated as more information unfolded throughout the study. This chapter provides a careful examination of the significant findings of the research in light of the little knowledge that currently exists for the sensitivity of error in offender risk classification. Also, this discussion addresses implications of the research findings, limitations of the study, and recommendations for further research.

Overview of Research Findings

Sensitivity and Classification Validity

Without having any prior knowledge on the sensitivity problem, the first question asks about the general impact of error on risk classification outcomes. In other words, how sensitive are risk assessment instruments? The study found that while many factors affect the sensitivity of error most risk assessment instruments could tolerate high levels of error. Specifically, the study found that the injection of 10 and 20 percent random and systematic error caused large levels of movement. Moreover, when looking at the subgroup base-rate changes subsequent the injections of such errors, the rates were significantly impacted at a statistical level of .05 or lower. However, turning to the comparison of subgroup base-rates, which Baird (2009) and Gottfredson and Snyder (2005) consider the best measure of validity, the injection of such errors did not significantly undermine the validity of the risk assessment instruments. Put differently, the risk assessment instruments continued to effectively divide individuals into groups with meaningfully different subgroup base-rates. Thus, risk assessment instruments generally are not very sensitive to error.

Research question #1 also distinguishes systematic error and random error by separately examining their impact on classification outcomes. Briefly, random error refers to the random distortion of information, where the displacement of cases is directionless. Conversely, systematic error refers to the distortion of cases that repeatedly goes towards one direction, either to the right or left. The study found that, under most circumstances, random error causes more case displacements than would systematic error.

There is one situation, however, that would reverse this relationship. When a high-leveled skew in cases is matched with a particular type of systematic error, an interactive effect may take place to magnify the impact of error. This was seen when upward systematic error was injected into each of the 30 risk items in the Oregon Instrument, which coincidentally contained very skewed items. Thus, the impact of systematic and random error could be mitigated or aggravated by the distributions. In other scenarios, however, random error produces greater levels of misclassification than would systematic error.

After understanding the general sensitivity of risk instruments and the differential impact of systematic and random error, the next step was to inquire about the distribution of cases across categories and its impact on sensitivity. The majority of cases can fall under: low risk; moderate risk; or high risk. Hypothesis #5 seeks to understand how the distribution of cases in the entire risk instrument affects the sensitivity of error. It was speculated that the skews of cases for an entire risk device would differentially impact the sensitivity of error. Specifically, negative skews in the cases of risk categories were hypothesized to have the gravest impact on increasing the sensitivity of error. However, upon further analysis, it was learned that positive distributions are most sensitive to error.

Hypothesis #6 tested the impact of increasing the number of risk categories in risk tools. The study found that increases to the number of risk categories also increase the number of misclassified cases. The manner in which risk device designers choose the number of categories for dividing individuals are sometimes arbitrary. It makes logical sense when a correctional facility has enough resources to deal with multiple levels of risky offenders. If so, a risk instrument with corresponding risk levels would be ideal, as

it would compliment the department's goals and budget. However, Baird (2009) argues that under most correctional circumstances, fiscal budgets are restrictive, and the facilities don't have sufficient resources to allocate to the individuals who fall into different risk designations. In this sense, the dividing of individuals as a function of most risk designs is capricious. Hypothesis #9 seeks to analyze this problem from a sensitivity perspective. The results fail to reject hypothesis #9; they also lend validity to the notion that increasing the number of risk categories will increase the sensitivity of error. The findings give additional support to the argument that unnecessary partitioning of individuals into more risk categories reduces efficiency.

Question 3 was comprised of 4 hypotheses, each examining the impact of different types of variables on the overall sensitivity of risk classification instruments. Hypothesis #7, which posited that negatively skewed risk items would have the greatest impact in increasing sensitivity, was confirmed. Item A from Risk Device X created the largest net change in subgroup base-rates, far greater than those from normally skewed (Item B) and positively skewed (Item D) risk items. Findings from the testing of hypothesis #7 much attention should be paid towards the skew of individual variables.

Next, risk variables could differ on their kurtosis. Hypothesis #8 was confirmed. Risk items that were more peaked displaced significantly more individuals to a different risk designation. Furthermore, the kurtosis seems to dictate the size and magnitude of the direction of displaced cases. In other words, having more peaked variables does not impact the quantity of misclassifications, rather it serves as an indicator of the magnitude of shifts towards a particular direction.

Dichotomous risk items are increasingly more popular in current risk instruments, and for good reasons. Reliability studies show that there is much higher inter-rater reliability when risk items are simple (Austin et al., 2003). Dichotomous risk variables minimize the discretion necessitated in the staff, thereby increasing reliability. In addition, such simple risk items make administering of risk assessments easier, more convenient, and manageable. Hypothesis #9 seeks to evaluate whether the inclusion of dichotomous variables is judicious from a sensitivity perspective. The hypothesis, which argues that dichotomous risk variables will have the greatest impact in increasing sensitivity of error, was rejected. In fact, the opposite relationship was discovered. Compared to risk items with three or more categories, dichotomous risk items displaced significantly less cases to other risk designations. Thus, dichotomous risk items reduce the sensitivity of error. Risk device designers are recommended to employ dichotomous risk items over other risk items with more options.

Hypothesis #10 argues that the weight of a risk item, relative to those of other risk items, have a direct and positive relationship to sensitivity. This supposition was supported. It is not uncommon to see certain risk devices include a variety of risk items that differentially contribute to the risk function. For instance, the LSI-R heavily relies on dichotomous risk items, but it also includes several risk items that contain a range of options that follow the score point schematic of 0, 1, 2, 3. Such items inadvertently alter the sensitivity of error, where the injection of error into them would produce greater levels of misclassifications. Thus, for the strict argument of ameliorating sensitivity, such items should be obviated and excluded from the risk function.

There has been an ongoing debate about the types of risk items that should go into a risk assessment instrument. For instance, the creators of the LSI argue that both needs and risk variables are both essential to the design of efficient risk assessment instruments (Andrews et al., 2006). One unintended side effect of this is that such risk/needs assessment instruments tend to contain many more variables. The LSI-R, for example, contains 54 risk/needs items in a single risk assessment instrument. On the other hand, Baird (2009) offers empirically supported arguments that the inclusion of needs variables or variables that do not explain enough variance in a risk function produces noise, undermining the efficiency of the instruments. Thus, following Baird's (2009) recommendation to remove unnecessary variables, risk devices should contain no more than six to eight essential risk items. Hypothesis #11 seeks to reconcile these opposing views by offering an argument about sensitivity. The sensitivity analyses offer support for the design of risk instruments with larger varieties of risk items. That is, as the number of risk items increases, the range of score will expand, offering a buffer for error. Thus hypothesis #11, which argues that the number of risk items will have a negative relationship with sensitivity, is supported.

The number of scores for each category is determined by the cut-offs. Increases in a category's range of scores should reduce the misplacement of individuals caused by error because there would be more room for error. This means that each score point difference would have a lower likelihood of causing classification error.

The range of scores refers to the spectrum of score points that comprise a risk function. The range of scores is affected by a multitude of variables such as: number of risk items; weight of items; risk item options; and inter-correlations of risk items.

Hypothesis #12, which states that a negative relationship exists between the range of scores and sensitivity, is supported. Like most other risk characteristics discussed, risk device designers are given a broad range of discretion and decision making power to select properties he/she sees suitable. It is hoped that the findings from the sensitivity analyses could provide designers with increased levels of guidance due to their increased knowledge of the problem of sensitivity.

This dissertation's findings suggest that more research should focus on the sensitivity of error in risk assessment instruments. Being that this is the first study to explore the impact of error on classification outcomes, the findings reported only represent the tip of an iceberg. Additional research and further in-depth analyses in risk assessment and sensitivity will greatly benefit risk devices. Specifically, risk device designers will gain a better understanding of the risk device properties that would moderate or aggravate the transfer of error.

Limitations

Especially in the world of social sciences, there exist real limits that could plague a study, often reducing or completely decimating the validity of findings when proper precautions aren't carefully considered, planned, and executed. The key, however, is to comb through the research methodology for such weaknesses, and to either preemptively remove the identified culprit or use this knowledge to guide the interpretations of the findings. The current study has three relevant limitations that are thought to have a real impact on the validity of the study. First, the selected methodology used to generate the data does not allow the user to manipulate changes in the correlations between risk items.

Second, there are real limitations in replicating correlated error. Third, the distribution of failed cases that are linked to the base-rate is unrealistic. Together, these three limitations threaten the internal and external validity of the study in ways that may undermine the findings.

Looking in from a purely statistical orientation, it would be ideal if individual risk items were independent and uncorrelated. There are a multitude of consequences related to multicollinearity, including the corruption of coefficient estimates and matrix inversion (Fields, 2005). Since these risk items are translated into independent variables that belong to a risk function, typically a logistical regression equation, high levels of inter-correlation would weaken the validity of the statistical model. Thus, one of the gravest violations that could be committed against regression models is when little is done about multicollinearity. Conversely, the best-case scenario would be such that the variables are completely independent. But, despite a conscious effort to reduce the inter-correlations among variables during the instrument design phase, inter-correlated items irrevocably find its way into risk instruments in the real world.

Thus, from a statistical standpoint, inter-correlations in variables cause invalidity in regressions because they explain the same variance in the dependent variable. Unfortunately, risk classification instruments often contain risk items with high levels of multicollinearity that which creates another problem. The existence of inter-correlations in variables in the real world present a redoubtable threat to the validity of the current study because the data construction method employed does not allow the researcher to create correlated risk data. In other words, we know that inter-correlations exist in risk

items, but unfortunately for the study, there does not exist a viable way to simulate inter-correlated data.

The inability to create correlated data in the current study translates into a credible external validity threat. Because of the over reliance on simulations to test the hypotheses posed, careful steps were taken to ensure that risk data, risk instruments, and multiple other scenarios are representative of real situations. Unfortunately, the STATA program and simulation method used do not allow the researcher to manipulate correlation levels among the items. Thus, simulated risk items are minimally correlated, and are much more random than real data. The results that are generated from the study will arguably lack representativeness because the simulated data does not fully reflect most actual data.

Next, the inability to create correlated error could contribute to threats to internal validity. A primary reason for the heavy focus on correlations is because we understand that correlated risk items would have profound impacts. For instance, if two risk items were highly correlated, then error in the information for one risk item would translate into error for the correlated item as well. Now, if multicollinearity exists across all risk items, the consequence would be devastating because error in one item could have serious implications for all other risk items with which it is correlated. The current study, unfortunately, only examines the impact of systematic and random error, but neglects to measure the impact of correlated error. Systematic and random errors affect individual cases, and thereby assume little or no multicollinearity in the risk items. However, it is likely that correlated error is more prevalent in the real world. As such, the inability to create correlated error could pose a potential threat to the internal validity of the study.

Finally, the base-rate distribution is unrealistic. In an attempt to construct a perfect risk data, the distribution of failed cases were assigned to the sample in a manner that caused an unusually high correlation between the risk score and criterion. In other words, there were too many actual failures in the high-risk group, and too many non-failures in the low-risk group. Risk scores from real risk data are, unfortunately, more loosely correlated with the outcome variable. As a result, Risk Device X demonstrates an atypically and unrealistically high capacity to divide individuals into groups with varying subgroup recidivism rates.

Policy/Future Research Implications

The research findings of this dissertation suggest several courses of action for risk assessment and future research. First, risk device designers should be more cognizant of the sensitivity issue associated with their risk instruments. It is seen that the right combination of aggravating factors could have a profound and detrimental impact on the sensitivity of error, where low levels of error could yield disproportionately high levels of misclassifications. Currently, risk assessments are evaluated based on validity, cost-effectiveness, equity and reliability, but little attention is given to the sensitivity of error. The findings offer support that sensitivity is equally important and its impact should be routinely evaluated.

Second, it could be argued that the manner by which risk properties are chosen is often arbitrary. With the exception of some general guidelines that exist to limit full subjective judgment, risk designers are free to choose risk items/properties as they deem fit. Such guidelines normally pertain to ensuring validity, reliability, equity, and cost effectiveness (Baird, 2009). However, the sensitivity problem is routinely neglected. In fact, a review of the past evaluations of risk assessment instruments indicates that no such research on sensitivity exists, and the focus of such evaluations typically focus on validity, reliability, and cost effectiveness. Thus, the understanding of how error is transferred through a risk device could help give risk designers some additional standard and criteria, by which the goodness of a risk instrument could be measured.

Third, using a classification system with a high sensitivity to error can substantially increase misclassification of individuals. By understanding specifically how

risk device properties affect transferred error and misclassifications, designers can better construct risk devices that are more tolerant of error. Also, by increasing the tolerance of error in risk classification, we reduce the need to frequently revalidate a risk device, which will save time and state resources.

The dissertation adds on to classification research that has not yet been directly studied. The knowledge we gain from the dissertation helps understand the way risk classification devices are designed. As a result, classification instrument designers will benefit by being able to remove properties that increase error and include properties that will reduce the transfer of error. Moreover, the criminal justice system will directly benefit from having more effective classification systems after knowing which specific factors will increase the sensitivity of error in offender risk classification instruments. The dissertation will make a true contribution to offender risk classification.

Finally, the findings of this study have a number of important implications for future research. The majority of the datasets was constructed using Monte Carlo Simulations. There are both strengths and weaknesses to this methodology, which was discussed earlier. However, research in classification outcomes and sensitivity would be better enhanced if such study received more access to different actual risk instruments and risk data. This would further allow for a better understanding of the different risk properties that differentially impact sensitivity, and by which risk instruments are set apart.

Next, research in sensitivity and risk classification should explore the different facets of correlated data, and how to effectively replicate it. Little understanding of the

impact of inter-correlated risk items on sensitivity was learned from the current study because it was unable to simulate correlated data and correlated error. Thus, newer and more effective methods of data simulation would allow us to construct data that are more reflective of real data. Future research should attempt to confirm the current study's theoretical findings by testing them on a wider range of datasets, whether simulated using alternative methods or borrowed from correctional agencies.

Reference

Alexander, J. & Austin, J. (1992). Handbook for evaluating objective prison classification systems. *National Institute of Corrections.*

Andrews, D.A. (1982). The Level of supervisory inventory (LSI): The First follow-up. *Ministry of Correctional Services.*

Andrews, D.A. (1982). The supervision of offenders: Identifying and gaining control over the factors which make a difference. *Ministry of the Solicitor General of Canada.*

Andrews, D.A. & Bonta, J. (2006). *The psychology of criminal conduct* (3rd ed.). Cincinnati, OH: Anderson Publishing Co.

Andrews, D.A., Bonta, J., & Wormith, J.S. (2006). The Recent past and near future of risk and/or need assessment. *Crime and Delinquency, (52)*1, 7-27.

Andrews, D.A. & Dowden, C. (2006). Risk principle of case classification in correctional treatment: A Meta-analytic investigation. *International Journal of Offender Therapy and Comparative Criminology, (50)*1, 88-100.

Andrews, D.A. & Robinson, D. (1984). The Level of supervision inventory: Second report. Toronto: Ontario Ministry of Correctional Services.

Auerhahn, D. (1999). Selective incapacitation and the problem of prediction. *Criminology, (37)*4, 703-734.

Austin, J. (2003). Findings in prison classification and risk assessment. *National Institute of Corrections: Prisons Division-Issues in Brief.*

Austin, J., Coleman, D., Peyton, J., Johnson, K.D. (2003). Reliability and validity study of the LSI-R risk assessment instrument. *Final Report: Submitted to The Pennsylvania Board of Probation and Parole.*

Austin, J., Hardyman, P.L., & Brown, S.D. (2001). Critical Issues and developments in prison classification. *National Institute of Corrections: Prisons Division-Issues in Brief.*

Baird, C. (2009). A Question of evidence: A Critique of risk assessment models used in the justice system. *National Council on Crime and Delinquency* 1-12.

Baird, C. (1991). Validating risk assessment instruments used in community corrections. *National Council on Crime and Delinquency.*

Baird, C. & Austin, J. (1987). Current state of the art in prison classification models. *National Council on Crime and Delinquency.*

Baird, C., Healy, T., Bogie, A., Danker, E.W., Scharenbroch,. C., Johnson, K. (2012). *Risk and needs assessments in juvenile justice: A Comparison of widely available risk and needs assessment systems.* Unpublished Manuscript.

Baird, C. & Lerner, D. (ND) *A Survey of the use of guidelines and risk assessments by state parole boards.* Unpublished Manuscript.

Baird, C. & Neuenfeldt, D. (1990). The Client management classification system. *The National Council on Crime and Delinquency,* 1-7.

Baird, C. & Wagner, D. (2000). The Relative validity of actuarial- and consensus-based risk assessment systems. *Children and Youth Services Review, 22,* 839-871.

Barnoski, R. & Aos, S. (2003). Washington's offender accountability act: An analysis of the Department of Corrections' risk assessment. *Olympia: Washington State Institute of Public Policy*, 1-16.

Belfrage, H. (1998). Making risk predictions without and instrument: Three years' experience of the New Swedish law on mentally disordered offenders. *International Journal of Law and Psychiatry, (21)*1, 59-64.

Berk, R.A. & Leeuw, J.D. (1999). An Evaluation of California's inmate classification system using a generalized regression discontinuity design. *Journal of the American Statistical Association, (94)*448, 1045-1052.

Berk, R.A., Sherman, L., Barnes, G., Kurtz, E., and L. Ahlman (2009). Forecasting Murder within a Population of Probationers and Parolees:A High Stakes Application of Statistical Learning." *Journal of the Royal Statistical Society* (Series A) 172, part 1: 191-211.

Blumstein, A. (1983). Selective incapacitation as a means of crime control. *American Behavioral Scientist, (27)*1, 87-108.

Blumstein, A., & Wallman, J. (2006). *The Crime Drop in America, 2nd ed.* Cambridge: Cambridge University Press.

Bonstedt, M., & Geiser, S. (1979). Probation/parole level of supervision source book. Washington D.C.: National Institute of Corrections.

Bonta, J. (2002). Offender risk assessment; Guidelines for selection and use. *Criminal Justice Behavior, (29)4,* 355-379.

Bonta, J. & Andrews, D.A. (2007). Risk-need-responsivity model for offender assessment and rehabilitation. *Her Majesty the Queen in right of Canada.*

Bonta, J. Harman, W.G., Hann, R.G., & Cormier, R.B. (1996). The Prediction of recidivism among federally sentenced offenders: A Re-validation of the SIR scale. *Canadian Journal of Criminology,* 61-79.

Bonta, J., Law, M., & Hanson, K. (1998). The Prediction of criminal and violent recidivism among mentally disordered offenders: A Meta-analysis. *Psychological Bulletin, (123)*2, 123-142.

Bonta, J. & Montiuk, L. (1996). High-risk violent offenders in Canada. *Correctional Research and Development.*

Brennan, T. (1987). Classification: An Overview of selected methodological issues. *Crime and Justice, (9), (*201-248).

Brennan, T. (1987). Classification for control in jails and prisons. *Crime and Justice (9),* 323-366.

Campbell, M.A., Frensh, S., & Gendreau, P. (2007). *Assessing the untility of risk assessment tools and personality measures in the prediction of violent recidivism for adult offenders.* Ottawa, ON: Public Safety Canada.

Carroll, J.S., Wiener, R.L., Coates, D., Galegher, J., & Alibiro, J.J. (1982). Evaluation, Diagnosis, and prediction in parole decision making. *Law & Society Review, (17)*1, 199-228.

Champion, D. (1994). *Measuring offender risk: A Criminal justice sourcebook.* Greenwood Press: Westport, CT.

Clear, T.R. (1994). The Design and implementation of classification systems. *Federal Probation*, 59(2): 58-61.

Clear, T.R. (1988). Statistical prediction in corrections. *Research in Corrections (1)*1, 1-40.

Clear, T.R. & Gallagner, K. (1985). Classification devices in probation and parole supervision: An assessment of current methods. *Crime and Delinquency, 31*(3), 423-443.

Clements, C.B. (1985). Towards an objective approach to offender classification. *Law and Psychology Review, (9),* 45-55.

Cullen, F.T. & Gendreau, P. (2000). Assessing correctional rehabilitation: policy, practice, and prospects. *In Criminal Justice 2000: Policies, Processes, and*

Decisions of the Criminal Justice System, ed. J Horney, 3:109-75. Washington, DC: U.S. Department of Justice, National Institute of Justice

Eisenberg, M. & Markley, G. (1987). Something works in community supervision. *Federal Probation, (51),* 28-32.

Erez, E. (1992). Dangerous men, evil women: Gender and parole decision-making. *Justice Quarterly, (9)*1, 105-126.

Farrington, D.P. (1987). Predicting individual crime rates. In D. Gottfredson & M. Tonry (EDs.) *Prediction and classification: Criminal justice decision-making.* (53-101). Chicago: University of Chicago Press.

Farrington, D.P. & Tarling, R. (1985). *Prediction in criminology*. State University of New York Press, Albany.

Feeley, M.M. & Simon, J. (1992). The New penology: Notes on the emerging strategy of corrections and its implication. *Criminology, (30)*4, 449-474.

Felson, M. (2002). *Crime and Everyday Life, Third edition*. Thousand Oaks, CA; Sage Publications.

Fields, A. (2005). *Discovering Statistics using SPSS.* London: Sage.

Flores, A.W., Lowenkamp, C.T., Smith, P. & Latessa, E.J. (2006). Validating the Level of Service Inventory-Revised on a sample of federal probationers. *Federal Probation, (70)*2, 44-48.

Flores, A.W., Travis, L.F., & Latessa, E.J. (2004). *Case classification for juvenile corrections: An assessment of the Youth Level of Service/ Case Management Inventory (YLS/CMI), executive summary* (98-JB-VX-0108). Washington, D.C.: U.S. Department of Justice.

Forst, B. (1984). Selective incapacitation: a Sheep in wolf's clothing? *Judicature, (68)4-5, 153-160.*

Gambrill, E. & Shlonsky, A. (2000). Risk assessment in context. *Children and Youth services review, 22,* 813-837.

Garson, G.D. (2003). *Statnotes: An online textbook.* Retrieved from http://www2.chass.ncsu.edu/garson/pa765/statnote.htm

Gendreau, P., Little, T., & Goggin, C. (1996). A meta-analysis of the predictors of adult offender recidivism: What works! *Criminology, (34)4, 575-607.*

Girard, L., & Wormith, J. (2004). The predictive validity of the Level of Service Inventory– Ontario Revision on general and violent recidivism among various offender groups. Criminal Justice and Behavior, 31, 150-181.

Glaser, D. (1987). Classification for risk. In D. Gottfredson & M. Tonry (EDs.) Prediction and classification: Criminal justice decision-making. Chicago: University of Chicago Press.

Glaze, L.E. & Bonzcar, T.P.(2006). Probation and parole in the United States, 2005. Washington, DC: Bureau of Justice Statistics, U.S. Department of Justice; 2006.

Glover, A.J., Nicholson, D.E., Hemmati, T., Bernfeld, G.A., & Quinsey, V.L. (2002). A Comparison of predictors of general and violent recidivism among high-risk federal offenders. *Criminal Justice and Behavior, (29)*3, 235-249.

Gottfredson, S.M. (1987). Prediction and classification in criminal justice decision making. In D. Gottfredson & M. Tonry (EDs.) *Prediction and classification: Criminal justice decision-making.* Chicago: University of Chicago Press.

Gottfredson, S.D. (1987). Prediction: An Overview of selected methodological issues. In D. Gottfredson & M. Tonry (EDs.) *Prediction and classification: Criminal justice decision-making.* Chicago: University of Chicago Press.

Gottfredson, D.M., Cosgrove, C.A., Wilkins, L.T., Wallerstein, J., & Rauh, C. (1978). Classification for parole decision policy. *National Institute of Law Enforcement and Criminal Justice.*

Gottfredson, S.D. & Gottfredson, D.M. (1994). Behavioral prediction and the problem of incapacitation. *Criminology, (32)*3, 441-474.

Gottfredson, S.D. & Moriarty, L.J. (2006). Statistical risk assessment: Old problems and new applications. *Crime and Delinquency, (52)*1, 178-200.

Gottfredson, D., & Snyder,H. (2005). *The mathematics of risk classification: Changing data into valid instruments for juvenile courts.* Washington, DC: US Department of Justice, Office of Justice Programs, Office of Juvenile Justice and delinquency Prevention.

Grant, B.A. & Luciani, F. (1998). Security classification using the custody rating scale. *Research Branch, Correctional Service of Canada.*

Haas, S.M. & Detardo-Bora, K.A. (2009). Inmate reentry and the utility of the LSI-R in case planning. *Corrections Compendium,* 11-16, 49-52.

Hagenbucher, G. (2003). Progress: An Enhanced supervision program for high-risk criminal offenders. *FBI Law Enforcement Bulletin, (72)*9, 20-24.

Hannah-Moffat, K. (2004). Losing ground: Gendered knowledge, parole risk, and responsibility. *Social Politics, (11)*3, 363-385.

Harcourt, B.E. (2007). *Against prediction: Profiling, policing, and punishing in actuarial age.* Chicago and London: The University of Chicago Press.

Harding, J. (2006). Some reflectins on risk assessment, parole and recall. *Probation Journal, (53)*4, 389-396.

Hardyman, P.L., Austin, J., Alexander, J., Johnson, K.D., & Tulloch, O.C. (2002). Internal prison classification systems: Case studies in their development and implementation. Washington DC: National Institute of Corrections

Hardyman, P.L., Austin, J., & Peyton, J. (2004). Prisoner intake systems: Assessing needs and classifying prisoners. Washington DC: National Institute of Corrections.

Harer, M.D. & Langan, N.P. (2001). Gender differences in predictors of prison violence: assessing the predictive validity of a risk classification system. *Crime Delinquency, (47)*4, 513-536.

Harris, P.M. (2006). What community supervision officers need to know about actuarial risk assessment and clinical judgment. *Federal Probation, (70)*2, 8-14.

Heibrun, K. (1997). Prediction versus management models relevant to risk assessment; The Importance of legal decision-making context. *Law and Human Behavior, (21)*4, 247-359.

Holsinger, A.M., Lowenkamp, c.T., & Latessa, E.J. (2003). Ethnicity, gender, and the Level of Service Inventory-Revised. *Journal of Criminal Justice, 31,* 309-320.

Jennings, W.G. (2006). Revisiting prediction models in policing: Identifying high-risk offenders. *American Journal of Criminal Justice, (31)*1, 35-50.

Josi, D.A. & Sechrest, D.K. (1999). A Pragmatic approach to parole aftercare: Evaluation of a community reintegration program for high-risk youthful offenders. *Justice Quarterly, (16)*1, 51-80.

Jones, D.A., Johnson, s., Latessa, E.J., & Travis, L.F. (1999). Case classification in community corrections: preliminary findings from a national survey. *Topics in Community Corrections.* Washington, DC: national Institute of Corrections.

Kassebaum, G., Davidson-Coronado, J. (2001). Parole decision making in Hawaii: Setting minimum terms, approving release, deciding on revocation, and predicting success and failure on parole. *Social Science Research Institute and Research and Statistics Branch Crime Prevention and Justice Assistance Division Department of the Attorney General.*

Kelman, M., Rottenstreich, Y., & Tversky, A. (1996). Context-dependence in legal decision making. *Journal of Legal Studies, (25),* 287-318.

Kemshall, H. (2002). Risk assessment and management of serious violent and sexual offenders: A Review of current issues. *Scottish Executive Social Research.*

Kennedy, P. (2008). *A guide to econometrics*. Massachusetts: Blackwell Publishing.

Kratocoski, P.C. (1985). The Functions of classification models in probation and parole: Control or treatment-rehabilitation. *Federal Probation,* 51-56.

Kubrin, C.E. & Stewart, E.A. (2006). Predicting who reoffends: The neglected role of neighborhood context in recidivism studies. *Criminology, (44)1,* 165-197.

Lowenkamp, C.T. Lemke, R. & Latessa, E. (2008). The Development and validation of a pretrial screening tool. *Federal Probation, (72)*3, 2-9.

Lowenkamp, C.T., Holsinger, A.M. & Latessa, E.J. (2001). Risk/Need assessment, offender classification, and the role of childhood abuse. *Criminal Justice and Behavior, (28)*5, 543-563.

Loza, W. & Loza-Fanous, A. (2001). The Effectiveness of the self-appraisal questionnaire in predicting offenders' postrelease outcome: A Comparison study. *Criminal Justice and Behavior, (28)*1, 105-121.

Maltz, M.D. (1984). *Recidivism.* Orlando Florida: Academic press.

Manchak, S., Skeem, J., & Douglas, K. (2008). Utility of the Revised Level of Service Inventory (LSI-R) in predicting recidivism after long-term incarceration. Law and Human Behavior, 32, 477-488.

MacKenzie, D.L. (2000). Evidence-based corrections: Identifying what works. *Crime & Delinquency, (46)*4, 457-471.

MacKenzie, D.L. (2006). What works in corrections? Reducing the criminal activities of offenders and delinquents. New York: Cambridge University Press.

Marutto, P. & Hannah-Moffat, K. (2006). Assembling risk and the restructuring of penal control. *The British Journal of Criminology, (46)*3, 438-454.

Mathiesen, T. (1998). Selective incapacitation revisited. *Law and Human Behavior, (22)4, 455-469.*

Meehl, P. (1954). *Clinical vs. statistical prediction: A theoretical analysis and a review of the evidence.* Minneapolis, University of Minnesota Press.

Miller, J. & Lin, J. (2007). Applying a generic juvenile risk assessment instrument to a local context: Some practical and theoretical lessons. *Crime and Delinquency, (53)*4, 552-580.

Montiuk, L. (ND). Classification for correctional programming: the Offender Intake Assessment process. *Research Branch, Correctional Service of Canada.*

Montiuk, L. & Porporino, F.J. (1989). Offender risk/needs assessment: A Study of conditional releases. *Research Report No. R-01: Research and Statistics Branch, Correctional Service of Canada.*

Onifade, E., Davidson, W., Campbell, C., Turke, G., Malinowski,J., & Turner, K. (2008, April). Predicting recidivism in probationers with the Youth Level of Service/Case Management Inventory (YLS/CMI). Criminal Justice and Behavior, 35(4), 474–483.

Orbis Partners. (2008). *Youth Assessment and Screening Instrument.* Retrieved on January, 6 2011, from www.orbispartners.com.

Petersilia, J. & Turner, S. (1987). Guideline-based justice: Prediction and racial minorities. In D. Gottfredson & M. Tonry (EDs.), *Prediction and classification: Criminal justice decision-making.* Chicago: University of Chicago Press.

Petersilia, J. 1999. "A Decade of Experimenting with Intermediate Sanctions: What Have We Learned?" In Perspectives on Crime and Justice. Washington, DC: National Institute of Justice.

Pressner, L. & Lowenkamp, C.T. (1999). Restorative justice and offender screening. *Journal of Criminal Justice, (27)*4, 333-343.

Rettinger, L.J. (1998). *A recidivism follow-up study investigating risk and need within a sample of provincially sentenced women* (Unpublished doctoral dissertation). Carlton University Ottawa, Canada.

Rhodes, L.A. (2004). *Total Confinement: Madness and Reasons in the Maximum Security Prison.* California: University of California Press.

Ruscio, J. (1998). Information integration in child welfare cases; An introduction to statistical decision-making. *Child Maltreatment, 3,* 143-156.

Sawyer, J. (1966). Measurement and prediction, clinical and statistical. *Psychological Bulletin, (66)*3, 178-200.

Sechrest, L. (1987) Classification for treatment. In D. Gottfredson & M. Tonry (EDs.), *Prediction and classification: Criminal justice decision-making.* Chicago: University of Chicago Press.

Serin, R.C., Peters, R.D., & Barbaree, H.E. (1990). Predictors of psychopathy and release outcome in a criminal population. *Psychological Assessment; A Journal of Counsulting and Clinical Psychology, (2)*4, 419-422.

Sherman, L. (1997). Policing for crime prevention. In University of Maryland, Department of Criminology and Criminal Justice (Eds.), *Preventing crime; What works, what doesn't, what'promising* (pp.8-1-8-58). Washington, DC; Office of Justice Programs, U.S. Department of Justice.

Silver, E. & Miller, L.L. (2002). A Cautionary note on the use of actuarial risk assessment tools for social control. *Crime Delinquency, (48)1, 138-161.*

Silver, E. & Banks, S. (1998). Calibrating the potency of violence risk classification models: The dispersion index for risk (DIFR). Washington, D.C.: American Society of Criminology.

Simourd, D. (2004). Use of dynamic risk/need assessment instruments among long-term incarcerated offenders. *Criminal Justice and Behavior, (31)*3, 306-323.

Stewart, D. (2008). The Problems and needs of newly sentenced prisoners: Results from a national survey. *Ministry of Justice.*

Tolman, A.O. & Rotzien, A. (2007). Conducting risk evaluations for future violence; Ethical practice is possible. *Professional Psychology: Research and Practice, (38)*1, 71-79.

Tong, J., & Farrington, D.P. (2006). How effective is the 'reasoning and rehabilitation' programme in reducing reoffending? A meta-analysis of evaluations in four countries. *Psychology, Crime and Law (12)1,* 3-24.

Tonry, M. (1987). Prediction and classification: Legal and ethical issues. In D. Gottfredson & M. Tonry (EDs.), *Prediction and classification: Criminal justice decision-making.* Chicago: University of Chicago Press.

Travis, J. 2005. But They All Come Back: Facing the Challenges of Prisoner Reentry. Urban Institute Press. Washington, DC.

Van Voorhis, P. & Brown, K. (1996). *Risk classification in the 1990's.* Unpublished Manuscript.

VanVoorhis, P., Salisbury, E., Wright, E. & Bauman, A. (2008). *Achieving accurate pictures of risk and identifying gender responsive needs: Two new assessments for women offenders.* University of Cincinnati Center for Criminal Justice Research, National Institute of Corrections, Washington DC.

Vinestock, M. (1996). Risk assessment. "A word to the wise"? *Advances in psychiatric treatment, (2),* 3-10.

Vose, Brenda A. 2008. Assessing the predictive validity of the Level of Service Inventory- Revised: Recidivism among Iowa parolees and probationers. Unpublished doctoral dissertation, University of Cincinnati.

Vose, Brenda, Cullen, F.T. & Smith, P. (2008). The empirical status of the Level of Service Inventory. Federal Probation, 72:22–29.

Wasson, B.F. (1988). The Use of prediction methods in a county corrections system. *Research in Corrections, (1)*1, 41-46.

Wilkins, L.T. (1985). The politics of prediction. In Farrington, D. & Tarling, R. (eds.) *Prediction in Criminology.* New York: SUNY Press.

Winterfield, L., Coggeshall, M., & Harrell, A. (2003). Deveolopment of an empirically-based risk assessment instrument: Final report. *Urban Institute- Justice Policy Center.*

Wright, K.N., Clear, T.R., & Dickson, P. (1984). Universal Applicability of probation risk-assessment instruments. *Criminology, (22)*1, 113-134.

Young, D., Moline, K., Farell, J. & Bierie, D. (2006). Best implementation practices: Disseminating new assessment technologies in a juvenile justice agency. *Crime Delinquency, (52)*1, 135-158.

Zhang, S., Farabee, D., & Roberts, R. (2007). Study of Parole Reentry in California. A presentation at the 66th Semi-annual Meeting of the Association for Criminal Justice Research (California), October 11-12, Long Beach, CA.

Zweig, J. M., Phillips, S.D., & Lindberg, L.D. (2002). Predicting adolescent profiles of risk: Looking beyond demographics. *Society for Adolescent Medicine, (31),* 343-353.

Curriculum Vitae

1985	Born in Manhattan, New York
2003	High School Diploma from Tottenville High School, Staten Island, NY
2007	Bachelor of Arts in Forensic Psychology, John Jay College of Criminal Justice, New York, New York
2010	Master of Arts in Criminology, Rutgers University, Newark, NJ
2007-2010	Teaching Assistant, Rutgers University, Newark, NJ, School of Criminal Justice
2010-2012	Part-Time Lecturer, Rutgers University, Newark, NJ, School of Criminal Justice
2013	Doctor of Philosophy in Criminal Justice, Rutgers University, Newark, NJ