DYNAMIC CHANGES IN LEVEL OF SERVICE INVENTORY-REVISED (LSI-R) SCORES AND THE EFFECTS ON PREDICTION ACCURACY

by

Thomas K. Arnold

B.A., St. Cloud State University, St. Cloud, 1983

A Thesis

Submitted to the Graduate Faculty

of

St. Cloud State University

in Partial Fulfillment of the Requirements

for the Degree

Master of Science

St. Cloud, Minnesota

May, 2007

This thesis submitted by Thomas K. Arnold in partial fulfillment of the requirements for the Degree of Master of Science at St. Cloud State University is hereby approved by the final evaluation committee.

Chairperson

Dean
School of Graduate Studies

DYNAMIC CHANGES IN LEVEL OF SERVICE INVENTORY-REVISED (LSI-R) SCORES AND THE EFFECTS ON PREDICTION ACCURACY

Thomas K. Arnold

The Level of Service Inventory-Revised (LSI-R) is one of the more popular risk/needs instruments in corrections. The predictive accuracy of the LSI-R had not been verified with the offender population used in this study, and a predictive validity study was done, which indicated that the LSI-R was a valid risk predictor for this population of offenders. One of the properties of the LSI-R, suggested by its authors, is the ability to measure changes in offender risk level. Previous research had verified that the predictive validity of the LSI-R increased from assessment 1 to assessment 2, but no study had been done to measure the dynamic predictive validity of additional assessments to determine whether the assessments continued to increase in accuracy. An attempt was made to measure the differences in predictive validity between four subsequent LSI-R assessments, and the results did not turn out as expected. There appeared to be a significant improvement in predictive validity from assessment 1 to assessment 2 as in the previous studies, but the last three assessments appeared to be very close, or almost identical, in predictive validity to each other, depending upon the measure used. The fourth assessment did not appear to improve at all in predictive validity over the previous two LSI-R assessments by any measure. There was a noticeable regression to the mean for the scores on each subsequent assessment and the impact of this trend is unknown at this time. These results suggest that further research in this area is definitely warranted.

____
Month Year

Approved by Research Committee:

____
Richard Lawrence Chairperson

ACKNOWLEDGEMENTS

I would like to thank my family for putting up with my spending so much time working on this. Thanks also to Dr. Lawrence, Dr. Prout, and Dr. Mayhew for their help in preparing this thesis. A very special thank you goes to Christopher Lowenkamp, James Bonta, and Peter Raynor for helping make this a fun project.

I would like to acknowledge the fact that, while we call them "offenders" the people who are the subjects in this study are real people, with real problems. I have had the pleasure of getting to know many "offenders" in weekly 2-hour visits to the local County Jail and I know that many of them would like to change their lives if they could figure out how to do it. This research is dedicated to finding better ways to help them to become productive members of society.

A Poem from “The Elephant’s Child”

By Rudyard Kipling (1956/1892)

I keep six honest serving-men (They taught me all I knew);
Their names are What and Why and When, and How and Where and Who.
I send them over land and sea, I send them east and west;
But after they have worked for me, I give them all a rest.
I let them rest from nine till five, for I am busy then,
As well as breakfast, lunch, and tea, for they are hungry men.
But different folk have different views; I know a person small
She keeps ten million serving-men, who get no rest at all!
She sends em abroad on her own affairs, from the second she opens her eyes
One million Hows, Two million Wheres, and seven million Whys!

TABLE OF CONTENTS

LIST OF TABLES

Table Page

Table | Page

LIST OF FIGURES

Chapter I

INTRODUCTION AND PROBLEM STATEMENT

The Level of Service Inventory-Revised (LSI-R; Andrews & Bonta, 1995) has been shown to be a valid and reliable tool for assessing offender risk levels for many offender populations (Gendreau, Goggin, & Little, 1996; Gendreau, Goggin, & Smith, 2002; Lowenkamp & Latessa, 2001; Girard & Wormith, 2004). One study had mixed results for the LSI-R however (O, Keefe, Klebe, & Hromas, 1998), suggesting that the LSI-R may need to be tested with each offender population to determine the relationship between LSI-R scores and subsequent recidivism rates. This study examined the relationship between the LSI-R scores for a Minnesota sample of offenders and the relationship of the scores to subsequent probation violation rates.

Because the LSI-R is composed of both static and dynamic factors, LSI-R risk assessment scores may change over time. There is very little research on the dynamic properties of the LSI-R (Andrews & Robinson, 1984; Motiuk, 1991; Motiuk, Andrews & Bonta, 1990; Raynor, 2007) and Andrews and Bonta (2003) had suggested that there is a need for additional research to determine whether these dynamic changes in LSI-R scores are correlated with subsequent changes in recidivism rates. This study examined the relationship between dynamic changes in offender LSI-R scores over time and the subsequent probation violation rates.

Rationale for the LSI-R Validation Study

Researchers have known for some time that the use of risk assessment instruments with offenders in different locations can be problematic. Wright, Clear, and Dickson (1984) found that there were problems with the predictive validity of the Wisconsin Case Management Classification System (CMC; Baird, Heinz, & Bemus, 1979) when it was used with offenders in New York and Ohio, despite the fact that it was a valid risk prediction instrument with the offenders it had been tested with. Problems have also been found with the LSI (Andrews, 1982), the forerunner of the LSI-R. O'Keefe et al. studied two groups of offenders in Colorado, and found that the LSI performed adequately with one group of offenders and not the other. This study will attempt to validate the LSI-R with the offenders served by a Midwestern Community Corrections Department to determine whether it is a valid risk predictor.

Rationale for the Dynamic Risk Prediction Study

Dynamic risk prediction is based upon the theories developed by Andrews and Bonta (2003; 2006), who assert that offenders can and do change, and risk prediction instruments should measure the dynamic correlates of risk. The LSI and its successor, the LSI-R, were designed to measure dynamic changes in risk level. There have been few empirical studies comparing dynamic changes in LSI-R scores with the subsequent risk levels. Two early studies were done on the dynamic properties of the LSI by Andrews and Robinson (1984) and Motiuk (1991) in Canada and a later study was done by Raynor (2007) with the LSI-R in Great Britain.

Brown (2003) had several criticisms of the early LSI studies. The main concerns cited were that the studies were retrospective in nature, they were done on small samples of offenders (n=57; n=55), the samples were not drawn randomly from the populations, the samples were a small portion of the total population (10%; 11%), and the study designs only used test and retest LSI scores. Brown suggested that future studies should be prospective, with random samples, and look at more than two assessments. The study done by Raynor on the dynamic properties of the LSI-R was a prospective study with a random selection of offenders and a larger sample size (360). The Raynor study was limited because it only used a test and retest.

In order to address the need for further research in this area, this study will attempt to replicate previous studies done on the dynamic properties of the LSI-R to determine whether their conclusions are valid. The present study is limited since it is a retrospective study using non-random samples. The benefits of this study will be in the use of a larger retest sample (N=1,173) that is a larger portion of the population (37%), and the opportunity to analyze up to four waves of LSI-R assessments.

Hypotheses

This study will test two hypotheses,

1) The LSI-R is a valid risk predictor for offenders served by a Midwestern County Community Corrections Department.
2) LSI-R scores from subsequent assessments are more accurate predictors of risk level than LSI-R scores from previous assessments.

Chapter II

LITERATURE REVIEW

OFFENDER ASSESSMENT OVERVIEW

Offender supervision serves two main purposes in criminal justice, the classification of offenders so that the most dangerous offenders can be kept from harming the public, and the classification of offenders for proper treatment matching (Latessa & Allen, 2003). There are many methods used to assess offenders. Bonta (1996) describes four primary methods, clinical judgment, actuarial assessment, risk/needs assessment, and risk/needs/case management assessment.

Clinical Judgment

Bonta classifies clinical judgment as the first generation of risk assessment, and describes it as subjective assessment, professional judgment, intuition, or gut-level feelings. Clinical judgment involves collection of relevant information and an unstructured interview. The caseworker uses his or her professional judgment to determine the best course of action. Bonta believes that a weakness of clinical judgment is the freedom that the clinician has in deciding which information is relevant. This could lead to situations where clinical judgment is subject to personal biases, which might then lead to concerns regarding accountability and fairness.

Bonta also asserts that clinical judgment is not as accurate as other methods of risk prediction. Andrews, Bonta and Wormith (2006) report that clinical judgment has a low average prediction success rate (.12), which is about a fourth of that found with other risk prediction methods. Gottfredson and Moriarty (2006) point out that, while mechanical assessment is usually superior to clinical judgment, there are instances where clinical judgment should be used, such as in unique situations, or situations where there is no relevant assessment tool.

Actuarial Risk Assessment

Bonta classifies actuarial risk assessment the second generation of risk assessment. Hornell Hart (1923) discovered the science of actuarial risk assessment. Hart described how he made a secondary analysis of data collected by Warner (1923). Warner had been commissioned by the Massachusetts Department of Corrections to try to determine whether there was any useful information in detailed survey information on 300 offenders who had broken parole, 300 offenders who had not, and 80 offenders who had served their sentences. Warner had reported that there was no useful information to be found in the data due to the methods of collection and the difficulties in analysis. Hart reanalyzed the data using the statistical methods of Yule (1919) and Davenport (1914) and found that there were significant differences between the offenders who recidivated and those who didn't. He determined that the odds of these differences occurring by chance were less than one in a million. He tabulated the differences between recidivists and non-recidivists and published his results, along with a description of his methods.

Five years later, Burgess (1928), using Hart's methods, published a report on the use of statistics to analyze data collected on 3000 offenders released from three prisons in Illinois. Burgess reported finding 22 factors that appeared to be related to non-criminal behavior and suggested a scoring system, using 21 factors, where scoring consisted of adding one point for each non-offending factor that was present. He tabulated the total number of offenders by score level and compared the results with the violation rate for each score level. He found that offenders with high scores violated parole at a much lower rate than offenders with low scores.

Bonta reports that Glueck and Glueck (1950) were also pioneers in the art of statistical prediction. Glueck and Glueck studied 500 delinquent boys and 500 non-delinquent boys who lived in Massachusetts. They studied the two groups for quite some time and developed detailed actuarial tables comparing many different behavior attributes between the two groups. The work of the Gluecks is still used to guide research (Sampson & Laub, 1993; 2005).

The main criticism of actuarial risk assessments made by Bonta is that the risk factors generally measure historical information that changes very little over time. Because of their static nature, they are not very useful as a guide for treatment decisions. Another criticism is that actuarial assessments are atheoretical and simply use known risk factors without explaining why they work. This practice makes progress difficult (Bonta, 2006). Theoretical considerations aside, it is clear that some actuarial risk assessments are excellent predictors of criminal behavior (Andrews, Bonta & Wormith, 2006).

Risk/Needs Assessment

Bonta classifies risk/needs assessments as the third generation of offender assessment. The main difference between actuarial risk assessments and risk/needs assessments is the use of dynamic risk factors, called criminogenic needs, that change over time. A predecessor of the LSI, the Wisconsin Case Management Classification System (CMC; Baird, Heinz, & Bemus), was one of the first classification systems to include needs assessments. The CMC was developed in Wisconsin in 1975 and became a model system recommended by the National Institute of Corrections (Latessa & Allen, 2003). The CMC was an improvement over previous systems for measuring risk, but had problems with the needs component of the CMC due to a lack of a theoretical basis for the needs, and there was a lack of research into the relationship between the needs and criminal activity (Andrews & Bonta, 2003).

In order to address some of the issues with the CMC, the Level of Supervision Inventory (LSI; Andrews, 1982) was developed in Canada during the late 1970s with funding provided by the province of Ontario (Bonta & Motiuk, 1987). The LSI was revised as the Level of Service Inventory-Revised (LSI-R; Andrews & Bonta, 1995). The LSI-R is a structured interview with 54 yes/no response items that are scored as either a 1 or a 0. Low scores indicate a low probability of criminal activity and high scores indicate a high probability. The LSI-R has ten sub-scales, one static, Criminal History, and nine dynamic, or changeable, Education/Employment, Financial, Family/Marital, Accommodations, Leisure/Recreation, Companions, Alcohol/Drugs, Emotional/Personal, and Attitude, which are related to the primary factors associated with risk of criminal conduct. One of the advantages of the LSI-R over actuarial risk assessment instruments is the ability to guide treatment decisions (Andrews & Bonta, 2003). Andrews Bonta and Wormith (2006) found that the LSI-R is one of the more accurate risk prediction instruments.

Risk/Needs and Case Management

The next generation of assessment instruments, described by Bonta as the fourth generation of offender assessment, combines the features of risk/needs instruments with case management. According to Andrews, Bonta, and Wormith (2006), fourth generation (4G) assessments guide and follow the supervision process from offender intake through release. The 4G assessments assess risks, strengths, needs, and offender responsiveness, or responsivity, to treatment. They can be linked to service plans and service delivery and measured through intermediate outcomes. They are designed to maximize adherence principles of effective treatment and to provide information that can improve treatment outcome in the future. The most well known 4G systems are the Wisconsin Correctional Assessment and Intervention System (CAIS; information available at www.nccd-crc.org/need_main.html), the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS; Northpointe Institute for Public Management, 1996), the Offender Intake Assessment (OIA; Motiuk, 1997), and the Level of Service/Case Management Inventory (LS/CMI; Andrews, Bonta, & Wormith, 2004). Andrews, et al. report that the accuracy for the LS/CMI when measuring general recidivism is about .41 which is about 15% better than the accuracy for the LSI-R (.36).

RESEARCH RELATED TO THE CURRENT STUDY

Validating the LSI-R

The original LSI was developed in Canada and was tested by Andrews (1982) on 598 offenders under the care of the Ottawa department of probation and parole. Its stated purpose was to provide an indication of the level of supervision required for each offender. The user manual for the LSI-R has reframed the construct that the LSI-R was purported to measure as “the propensity for rule violations” (Andrews & Bonta, 1995). There have been many studies of the LSI-R, prompting Hollin (2002) to comment that the LSI-R had the strongest “research pedigree” of any risk prediction instrument. There are two meta-analyses comparing the LSI-R with other risk assessments (Gendreau, Little, & Goggin, 1996; Gendreau, Goggin, & Smith, 2002).

Gendreau, Little, and Goggin compared the results of 123 studies of risk scales, including 28 LSI-R studies, 15 studies of the Salient Factor Scale (SFS; Hoffman, 1983), 14 studies of the Wisconsin CMC, and 66 studies of other risk scales. They found that the mean effect size for the LSI-R was .35, which was higher than the SFS (.29), the Wisconsin (.27), and various other risk scales (.30). They also examined several measures of antisocial personality for their level of risk prediction and found that the LSI-R was a better predictor than the Psychopathy Checklist-Revised (PCL-R; Hare, 1980; 1991; 2003) which had a mean effect size of .28, the Minnesota Multiphasic Personality Inventory (MMPI; Megaree & Bohn, 1979) which had a mean effect size of .16, and 37 other personality scales which had a mean effect size of .16.

Gendreau, Goggin, and Smith compared the LSI-R with the PCL-R and found that the LSI-R had a mean effect size of .37 for general recidivism and .26 for violent recidivism, while the PCL-R had a mean effect size of .23 for general recidivism and .21 for violent recidivism.

From the results of these two meta-analyses, it is clear that the LSI-R is generally an effective risk prediction instrument, however that does not mean that the LSI-R performs equally in all locations. In the Appendix of the Gendreau, Goggin and Smith study, the list of the effect sizes for the studies they analyzed shows that there is a wide range of correlation rates that were found for the LSI-R with different offender populations. There would appear to be a need to study the LSI-R with each offender population if knowledge of the exact risk prediction level is desired.

There was one study done on the LSI-R in Minnesota (Jenson, 1998). The results found in that study indicate that the LSI-R was able to distinguish between low-risk offenders and medium or high-risk offenders, but not between the medium-risk and high-risk offenders. The study uses a small sample (n=66) and did not appear to have explained which score ranges made up the low, medium, and high levels.

Previous validity studies have used a variety of methods for validation. The original Andrews study looked at the total LSI scores and the inter-correlation rates between sub-scale scores. The LSI scores were divided into risk levels, low (0-7), medium (8-11), etc. and the failure rates were determined by risk level. The failure rates increased as the risk levels went up. The correlation rate between LSI scores and parole failure was calculated and was determined to be .40 for the LSI-VI revision.

<u>The Dynamic Predictive Validity of the LSI-R</u>

There have been few studies on the dynamic predictive validity of the LSI-R. The most notable studies, to date, were done by Andrews and Robinson (1984), Motiuk (1991), Motiuk, Bonta, and Andrews (1990), and Raynor (2007). The earlier studies used small samples of between 50 to 60 Canadian offenders, and the study by Raynor used a sample of 360 offenders in Great Britain.

Andrews and Robinson (1984) studied 57 offenders in Canada who had both a test and retest done with the original Level of Supervision Inventory (LSI) with a follow-up period of at least 18 months. They divided the offenders into four levels by LSI score. Offenders with a score of 0-7 were placed in the low risk category, LSI scores of 8-11 were moderate risk, LSI scores of 12-23 were high risk and LSI scores of 24+ were placed in the very-high risk category. They totaled the recidivism and outcome rates for offenders using the initial and retest LSI categories and compared the results in a table. Recidivism was measured by reports from probation officers and self-reports by offenders. The outcome criteria was based on a score from 0 to 2 in which early termination or closure without recidivism was coded as a 0, regular termination without recidivism was coded as a 1, and recidivism was coded as a 2.

Andrews and Robinson found that the offenders were more accurately placed in risk categories by the retest LSI than the initial LSI. They concluded that the retest LSI predicted recidivism and outcome better than the initial LSI. When they measured the outcome for the LSI subtotals using both the initial and retest values, they found that all of the retest LSI subtotals except Attitudes more strongly predicted recidivism than the initial LSI subtotals. The retest LSI subtotals most strongly linked with recidivism were Companions and Leisure/Recreation. The retest LSI subtotals with the weakest links to recidivism, besides Attitudes were the Family and Emotional/Personal subtotals.

Motiuk (1991) studied the post-release outcomes of 54 Canadian offenders given an LSI assessment upon intake to prison and then given a follow-up LSI before release. He used two LSI risk categories, low (0-19) and high (20-54). He found that when retest risk level increased to the high level from initially being low, post–release remands, incarceration, recidivism, and parole violation increased. When retest risk level decreased to the low level from initially being high, post–release remands, incarceration, recidivism, and parole violation decreased. When outcomes for violent recidivism, violent re-offense, and Federal sentence were compared with LSI intake/retest risk levels, the results were mixed. When LSI risk levels decreased from the first to second assessment, the levels of violent recidivism, violent re-offense, and Federal sentence decreased. When LSI risk levels increased, the incidence of violent recidivism, violent re-offense, and Federal sentence did not show an increase.

Motiuk, Bonta, and Andrews (1990) reported additional results from the Motiuk study related to the LSI-R sub-scales. When they analyzed the correlation rates between the intake LSI and retest LSI sub-scales and subsequent incarceration, they found that retest scores for Education/Employment, Accommodations, and Drug/Alcohol were more highly correlated with a negative outcome, while the initial scores for Financial, Family/Marital, Leisure, Companions, and Attitude sub-scales were more predictive than retest scores. The Emotional/Personal sub-scale was more predictive at retest of incarceration and more predictive at the initial test of recidivism. A regression analysis was done on the test and retest LSI scores to determine which test was better able to explain variance between the scores and outcome. It was found that the retest scores had a 107% percentage gain in explained variance (PGV) in predicting incarceration and a 64% PGV in predicting general recidivism.

Raynor studied 360 offenders in the British Isles who had follow-up assessments done with the LSI-R. Due to concerns with regression to the mean in the follow-up LSI-R scores he split two samples of offenders, one from England and Wales, and the other from Jersey, into increasing and decreasing categories for offenders whose scores on the first LSI-R were both above and below average. He compared the recidivism rates for offenders from both locations whose scores were below average and decreasing, below average and increasing, above average and decreasing, and above average and increasing. He found that for offenders from England and Wales with above average scores on the first assessment, when scores increased on the second assessment, there were significantly larger reconviction rates ($p<.01$) than for offenders with decreasing scores. For Jersey offenders with above average scores on the first assessment, offenders who had increasing scores on the second assessment had larger reconviction rates than offenders with decreasing scores, but the difference did not reach the level of significance ($p=.06$). He speculated that the lack of significance for the Jersey offenders might have been due to a small sample size of 21 offenders. Offenders from both groups of offenders who started with below average scores on the first assessment, had significantly higher ($p<.05$) reconviction rates if their scores on the second LSI-R were increasing than if their scores were decreasing. Combining all offenders with increasing LSI-R scores and comparing them with offenders with decreasing scores, showed offenders with increasing scores having higher reconviction rates (67%) than offenders with decreasing scores (42%).

The Raynor study was an improvement over previous LSI-R studies due to its prospective nature, the larger sample used, and the random sampling method. It also was an improvement in the fact that it compared the LSI-R with another risk instrument for comparison. It shared a concern with the other studies in that it only used a test and retest for analysis.

Other research on dynamic predictive validity. There are additional studies that have been done to look at the dynamic predictive validity of other risk assessment instruments. Brown (2003) did a thorough in-depth review of most, if not all, of the studies available at that time. Many of the studies shared a design flaw in that they only looked at two waves of assessment scores. The Brown study attempted to rectify that deficiency. Her conclusion was that the dynamic components of the risk assessment instruments she used, added significantly to the overall predictive accuracy of the assessments.

Another problem that was discussed by Brown is the conceptualization of change with regard to how fast change occurs. Brown notes that different facets of the offender's situation change at different rates, with some properties being very labile and changing in minutes and others possibly taking months or years to change.

ISSUES WITH THE CURRENT STUDY

There are two parts to this study, validating the LSI-R, and analyzing the dynamic predictive validity of the LSI-R. The first part of this study is fairly straightforward. The LSI-R is a summated measurement scale that has been validated as a risk prediction instrument in other locations, and there is a need to assess its predictive validity with the current population. Andrews (1982) has demonstrated the basic methods used to accomplish validation and so this part of the study will simply replicate earlier research.

The second part of this study, analyzing the dynamic predictive validity of the LSI-R, is not so simple. The problem of how to conceptualize dynamic risk prediction is not unique to research with the LSI-R. Brown (2003) notes that there is no standard in criminal justice as yet for how changes in risk should be measured. Since a conceptualization of the processes involved in measuring change is essential to this study, an attempt will be made to look at this topic from various perspectives. Due to the lack of discussion of this topic in criminal justice, the theory and research into the measurement of change from several other disciplines will be examined. The most promising places to start appear to be with the subjects of quasi-experimental research design, classical test theory, medical research and education research. While the LSI-R does not fit exactly into these domains, there are similarities. For instance, in education testing, there is a desire to look at multiple assessments and determine whether a measurable change is occurring in the student. In the case of offender research, multiple assessments are also given to measure change in the offender.

Single Wave Designs

Most LSI-R research has used a single wave design in which a single LSI-R assessment was done and then violations were measured for some time period. Using the notation of Cook and Campbell (1979) with two observations, the design diagram looks like this [O_{1A} O_{1B}]. O_{1A} is observation with the LSI-R and O_{1B} is observation by the police or probation department during the follow-up period. Both observations are attempting to measure the propensity for rule violation (PRV). The LSI-R is using observation of the offender's life situation to measure this construct and the police are using observation of behavior. Since the second observation happens some time after the first observation, the correlation rate between the two observations is a measure of predictive validity (Cronbach & Meehl, 1955).

True score theory, a.k.a. classical test theory, states that both measurements consist of a true value and an error value (Crocker & Algina, 1986). In theory, if both the LSI-R and the police were able to determine the true value of the PRV, the correlation rate between the measurements would be 1. The difference between the correlation rate and 1 is due to the errors of measurement. The possible causes of error could be low statistical power due to small sample size, poor reliability of the measures, variation in treatment of the offenders, random events that occur after the first observation, and random differences between offenders (Cook et al.). These possible causes of error could apply to both the LSI-R scores and the rate of rule violation measured by the police. The number of possible confounds is large and is difficult to estimate. The error rate is generally greater than the accuracy rate.

Multi-Wave Designs

When looking at multi-wave research, there is more than one observation made with the LSI-R. In a two-wave design the diagram becomes [O_{1A} O_{2A} O_{2B}] where O_{1A} and O_{2A} are the first and second LSI-R scores and O_{2B} is the observation by the police after the second LSI-R. Possible reasons given by Cook and Campbell for a difference between scores include historic events that occur between measurements, changes in the offender, errors in testing, changes in the measuring instrument, statistical regression a.k.a. regression to the mean, a loss of some of the test participants, and various interactions between the aforementioned possibilities. A more thorough treatment of these issues will be given in the section on individual issues, after a discussion of the measurement of change.

The Measurement of Change

When scores change, there are problems with deciding what to do with the new score. In the case of the LSI-R assessments, one could simply throw out the old score and use the new score. This is basically what Andrews and Robinson (1984) suggested in their study of test and retest scores. They compared the recidivism rates for both the initial LSI and the retest LSI scores to predict recidivism after the second LSI and found the second LSI scores to be better predictors.

If one wanted to use the difference between the scores, there are several issues to consider. If the changes in scores remain the same, did all of the individual items on the LSI-R remain the same, or did some change from 0 to 1 and others change from 1 to 0? If the scores changed, what are the magnitudes, directions, and errors in the changes? According to Crocker and Algina (1986), when looking at change scores, there is a true value of the change plus an error value of the change. If the error in the change score is non-zero and positive, the second score will have more error. If the error in the change score is negative, the second score will have less error. If the error in the change score is 0, the amount of error will be the same.

The issue with the error of changes in scores does not appear to have received much discussion in criminal justice, but has received considerable attention in the education field where the difference in test scores is desired in order to show progress between assessments. Cronbach and Furby (1970) highlighted some of the difficulties involved with trying to estimate the reliabilities of change scores. Rogosa, Brandt, and Zimowski (1982) continued the discussion and determined that there are ways to deal with the issue of the reliability and validity of change scores.

Willett (1989; 1994) argues that, while two wave data can be used, multi-wave research with three or more waves is preferred. In a discussion of the issues involved, Willett breaks the problem of assessing change into two parts, determining the individual changes and then looking at between person patterns of change. He goes on to discuss the relative merit of several different types of analysis.

Raudenbush (2001) reviews the current state of research in this area and concludes that issues regarding the measurement of longitudinal data are ongoing. Since this study was not intended to be the last word in this area, these issues will be left for future research. The issues involved with reassessment score changes will be discussed in terms of pretest and post-test outcomes using two waves of assessments.

Measurement and Testing Issues

Several possible issues were mentioned previously with regards to single wave and multi-wave research. A brief overview of the nature and scope of some of those issues, and others, will be given here.

Coefficient Alpha. One of the issues with single wave assessments is internal consistency within the test scores. The coefficient alpha is a statistical measurement that was developed by Cronbach (1951) and is widely used as a measure of reliability (DeVellis, 2003). DeVellis indicates that coefficient alpha is a measurement of the proportion of the score that is the true score. The value of coefficient alpha is $1 – e_v$ where e_v is the error variance. Bonta (1985) appears to be the first to have published reliability values for the LSI using the coefficient alpha.

Area under the curve (AUC). Correlation rates used in risk prediction are subject to errors due to low base rates of offending in the population (Meehl & Rosen, 1955). The Area Under the Curve (AUC), another method for measuring prediction accuracy, is a numerical representation of the area under a Receiver Operating Characteristic (ROC) curve. The ROC curve was initially developed to help improve signal detectability in radio waves (Peterson, Birdsall, & Fox, 1954), and has since been adapted for the measurement of diagnostic accuracy (Hanley & McNeil, 1982; Swets, 1973). ROC curves have been recommended for use in assessing risk prediction instruments because they are less sensitive to problems regarding base rates (Rice & Harris, 1995).

Regression to the mean (RTM). Regression to the mean (RTM) is an issue that seems to have been largely overlooked in criminal justice, although it appears to be an issue that is of concern in Great Britain (Raynor, 2007). A search for "regression to the mean" and "offender" in the criminal justice archive returned 27 articles. The term was generally used to discredit someone else's findings (French & Gendreau, 2006), as a reason for using a particular statistical method (Whaley, 2001), or was discounted as a possible cause for the results found (Mitchell & MacKenzie, 2006). RTM issues are discussed in medical research and James (1973) warns that regression to the mean is a problem in uncontrolled clinical studies where there is a test and posttest.

Stigler (1997) points out that RTM is rather ubiquitous and has lead many researchers to come to erroneous conclusions. RTM is one of the threats to internal validity mentioned by Cook and Campbell (1979). Cook and Campbell discuss how RTM, or statistical regression, results in high scores on the pretest being generally lower on the post-test, low scores on the pretest being generally higher on the post-test, and medium scores on the pretest remaining about the same on the post-test.

The discovery of RTM is generally attributed to Francis Galton (1886) who reported that children of unusually tall or unusually short parents tended to be more average in height. He also noted that when pea plants had extremely large or extremely small diameter peas, later generations of the plant had more average diameter peas. The problem of RTM in successive generations has been found to be a problem whenever multiple tests are given over a period of time. Galton also observed the opposite situation, the spreading of values between generations.

Raynor (2007) had suggested that RTM may be an issue with studies of the dynamic validity of the LSI-R and had proposed that the offender samples be split in two halves in order to minimize the effects of RTM. The recidivism rates for offenders with increasing and decreasing scores were then measured for both halves and compared to determine whether they had different rates of offending.

Changes in rater scoring between assessments. Two issues with rater scoring that can lead to differences in subsequent LSI-R scores, are changes within the rater between tests and changes due to different raters doing the test and retest. A study done by Flores, Lowenkamp, Holsinger, and Latessa (2006) showed that staff training practices and experience produce changes in rater scoring effectiveness. They looked at raters in several Midwestern Community Corrections departments and found that LSI-R raters with less than 3 years of experience had a .14 correlation relation rate between LSI-R scores and outcome and trainers with more than 3 years experience had a .25 correlation rate. There are also differences noted between raters. Lowenkamp, Holsinger, Brusman-Lovins, and Latessa (2004) looked at the percentage of change in LSI-R sub-scale scores when two different raters assess the same offender. The overall agreement between raters was around 90% for most of the LSI-R sub-scales, with the Financial sub-scale having the lowest agreement level.

Loss of participants. Cook and Campbell mention that shrinking sample size can be a threat the statistical conclusions in a test and retest design. The fact that the sample sizes in this study shrink substantially will be a cause for concern.

Rater experience. Andrews and Bonta (1995) discussed the results from the dynamic LSI-R study done by Andrews (1984). They suggested that changes in LSI-R scores between assessments could be due to actual changes in the offender's risk level or the could possibly be due to an improved understanding of the offenders gained by the LSI-R raters through additional experience in the time between assessments.

Reliability of measures. In research reports on the LSI-R, the reliability of the LSI-R is often mentioned, but little discussion is made of the inherent unreliability of some of the outcome measures that are used. It is highly likely that some offenders are breaking the rules and not getting caught. This would tend to undercount offenders who are rule violators. An estimate of the rate of offending over the arrest and conviction rate made by Blumstein and Cohen (1979) was that 9 to 17 index crimes per year per offender were committed while the average offender was free. The Minnesota BCA (2005) indicated that only 50% of the crimes committed in the county used for this study were cleared in 2005. These figures suggest that the outcome measure may have some variation due to the luck of the offender.

Under reporting violations may cause problems due to the non-normality of the errors in measurement, which are assumed to be equally likely in both directions (Crocker, et al.). Unless the police are arresting and convicting innocent people, the probability of errors in under reporting violations are greater than the probability of over reporting rule violation. Cook and Campbell mention the violation of statistical assumptions as a possible threat to the validity of statistical conclusions.

Chapter III

METHODS

Participants and Data Sources

The participants in this study consisted of the interviewers who conducted the LSI-R assessments and the respondents to the interviews. The respondents were offenders who were placed in the care of a Central Minnesota Community Corrections department. The data sources used in this study consisted of LSI-R records, arrest records, and parole violation records resulting in a commitment to prison.

Community Corrections LSI-R data. The LSI-R records used in this study were generated by probation officers from a Central Minnesota County Community Corrections Department and were obtained with the provision that all identifying characteristics be kept confidential. The records were a subset of data kept in a larger LSI-R database that was maintained by the State of Minnesota for the corrections departments in Minnesota that use the LSI-R. This subset contained all of the records created by the County Community Corrections Department for offenders placed in community corrections from 2002 through the latter part of 2006. The records were provided in a Microsoft Access Database format and included names, birthdates, LSI-R completion date, scores, sub-scale totals, and overall score totals.

The County provided data containing 8,860 separate LSI-R assessment records, including initial LSI-R results on 5,111 individual offenders, and at least one follow-up assessment on 1,866 offenders. The total number of follow-up LSI-R assessments was 3,749 with the number of assessments per offender varying in number from one to eight. Ten of the individual records were excluded from the data set because they were not completed, leaving 5,101 individuals for analysis.

Minnesota BCA data. The names and birthdates were exported from the LSI-R records to a Microsoft Excel file and sent to the State of Minnesota Bureau of Criminal Apprehension (BCA). The BCA matched 4,918 of the original 5,101 names and birthdates with offender violation and conviction records and returned the resulting data in a text file on a CD. The BCA records were provided with the provision that all records remain confidential with regards to individual characteristics.

There were 183 individual records (3.6% of offenders) that could not be matched with BCA data. The missing records tended to be significantly more female (40% vs. 20%), have lower LSI-R Scores, (15.9 vs. 22), and fewer prior violations (2.4 vs. 4.2). The race of these offenders is unknown since the racial characteristics were collected from the BCA database. The mean ages of 31.3 for the missing records and 32.4 for the matched records were not significantly different (p=.184).

Court services data. The County provided commitment to prison dates for all offenders who violated the terms of their probation and were sent to prison to serve their sentence. This data was also provided under a confidentiality agreement.

Data modification. There were three modifications made to the original data for ease of computation. 1) Null fields in the scoring fields were changed to 0 in order to prevent program errors. This was not seen to be a major issue as they had not been included in the totals anyway. 2) About 20 birthdates that were not coded correctly in the original data were collected from the State BCA database by manually matching offender information using the offender names, location of violation, LSI-R completion date, and violation date for identification purposes. 3) The incomplete records that were either unfinished LSI-R records or unmatched BCA records were deleted from the working table after the initial demographic information was collected.

Sample selection. To avoid undercounting violations, only data from assessments done from 2002 through 2004 were used in the study phase. This allowed for the analysis of 12-month recidivism rates with an additional year from the end of the study period for any violations to turn into convictions and be entered into the BCA database. It is assumed that most violations come to trial and turn into convictions within 1 year after the violation date. Sample 1 consisted of the LSI-R records of all of the offenders with an assessment completed before 2005. Samples 2, 3, and 4 consisted of the follow-up assessment records of offenders who received a second, third, or fourth assessment before 2005. Sample 2 is the subset of Sample 1 with two assessments, Sample 3 is the subset of Sample 2 with three assessments, and Sample 4 is the subset of Sample 3 with four assessments.

Inter-item reliability. Reliability calculations were performed on the 54 items of the LSI-R assessments to determine the Cronbach's alpha score for each sample and placed in Table 1. The overall inter-item reliability was high. The Cronbach's alpha for individual raters was between .78 and .92 except for one rater who had an alpha of .445 for 10 assessments. Crocker and Algina (1986) state that the coefficient alpha is a convenient way to estimate the lower bound of the precision of a test with data obtained from a single administration of the test.

Table 1

Sample Counts and Reliability Data

Sample	**N**	**Cronbach's alpha**
Sample 1	3190	.90
Sample 2	1173	.87
Sample 3	616	.88
Sample 4	285	.88

Simplification of data display. Previous research by Andrews (1982) had divided the LSI-R scores into risk levels in order to simplify the display. Initial analysis of the offender LSI-R scores in the population suggested that five roughly equal (20%) categories could be obtained by using 0-11, 12-18, 19-24, 25-31, and 32-54 as the score ranges. The offenders in Samples 2, 3, and 4 are under represented in the lower score categories and over represented in the higher score categories when compared with Sample 1. A side-by-side comparison of the percentages of offenders in each score category for the four samples is shown in Figure 1.

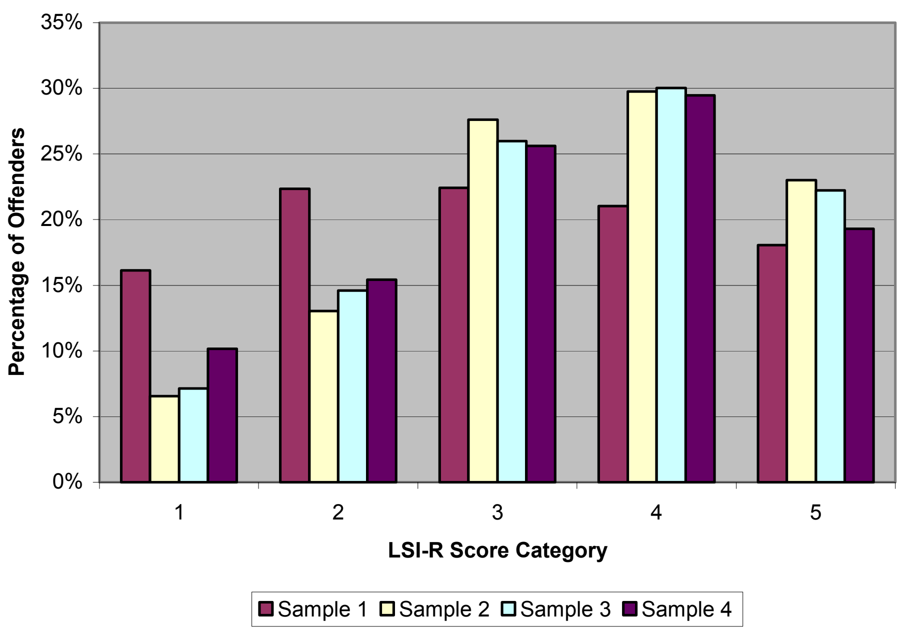

Figure 1

Comparison of Percentages of Offenders per LSI-R Score Category in Samples 1, 2, 3 and 4

Sample 1 demographic data. The demographic information for Sample 1 was compiled and is compared to the demographic information of the 4,918 records matched with the BCA in Table 2. A t-test was done on the age and the Mann-Whitney U test was used on the categorical values to determine whether the sample statistics for the various demographic items were significantly different from the values of the remaining 1,728 offenders who were assessed in 2005 and 2006. The only significant difference was the gender mix. Sample 1 had 81% male offenders and the excluded offenders were 77% male ($p<.01$).

Table 2

Demographic Information for Offenders Matched with the BCA, Offenders in Sample 1, and Offenders Not Included in Sample 1

	All Records	**Sample 1**	**Records Not Included**	**p**
N	4918	3190	1728	
Mean Age	34.57 S.D.=10.77	32.57 S.D.=10.84	32.09 S.D.=10.92	.145[a]
Male	80%	81%	77%	.003[a]
Female	20%	19%	23%	.003[a]
Race				
White	82% N=4032	82% N=2604	83% N=1428	.380[b]
Black	13% N=624	13% N=410	12% N=214	.638[b]
Native	3% N=137	3% N=96	2% N=41	.195[b]
Asian	2% N=91	2% N=62	2% N=29	.510[b]
Unknown	1% N=34	1% N=18	1% N=16	.144[b]

Note: [a] t-test probability of no difference in mean, [b] Mann Whitney U probability of no difference

Sample 2 demographic data. The demographic makeup of Sample 2 was calculated by score category and is shown in Table 3. The offenders with the lowest scores appear to be older, male, and White or Asian. The ratio of Black offenders to White offenders shifts dramatically as the score level increases.

Table 3

Demographic Breakdown by LSI-R Score Category for Sample 2

			LSI-R Score Category			
	0-11	**12-18**	**19-24**	**35-31**	**32-54**	**Total**
N	84	242	328	300	219	1173
Age	37.39	37.27	34.00	32.71	32.18	34.25
Male	90%	83%	82%	84%	83%	83%
Female	10%	17%	18%	16%	17%	17%
Race						
White	88%	92%	82%	78%	67%	81%
Black	1%	7%	13%	17%	24%	14%
Native	1%	1%	3%	4%	5%	3%
Asian	10%	0%	2%	1%	4%	2%
Unknown	-	-	-	0% (1)	-	0%

Data Variables

The independent variables used in this study were age at assessment completion, gender, race, LSI-R item scores, total LSI-R score, rater, completion date, and LSI-R sub-scale scores. The dependent variable used in this study consisted of the occurrence of a parole violation, either in the form of a subsequent arrest resulting in conviction, or a technical violation that resulted in a commitment to prison.

<u>Data coding.</u> Age was measured in years and consisted of the number of years from birth date to LSI-R completion date. Age was not rounded upwards. Gender was coded as a 1 for male and a 0 for female. Race was coded as either White or Non-White with a 1 indicating White and a 0 indicating non-white. Individual LSI-R items were coded as a 1 or a 0 with 1 representing a yes and 0 representing a no. The LSI-R score was an integer with possible values from 0 through 54 that represented the total number of yes answers on the LSI-R assessment. Rater id was a nominal representation of the rater and was used to remove all identifying characteristics. The LSI-R completion date was coded as a date. The LSI-R sub-scales were used in the validation portion of this study. The LSI-R sub-scales, question numbers, and number of items per sub-scale are listed below in Table 4 (Andrews, 1982).

Table 4

LSI-R Sub-Scales, Question Numbers, and Number of Items per Sub-scale

LSI-R Sub-Scale	**Question Numbers**	**# of Items**	**LSI-R Sub-Scale**	**Question Numbers**	**# of Items**
1. Criminal History	1-10	10	6. Leisure and Recreation	30-31	2
2. Education/Employment	11-20	10	7. Companions	32-36	5
3. Financial	21-22	2	8. Alcohol and Drugs	37-45	9
4. Family/Marital	23-26	4	9. Emotional and Personal	46-50	5
Accommodations	27-29	3	10. Attitude and Orientation	51-54	4

The dependent variables were coded as a 1 or 0, with a 1 indicating a probation violation after completion of the LSI-R assessment. Two separate dependent variables were used, probation violation within 6 months and violation within 1 year.

Equipment

A fast Pentium computer with 2G of RAM was used because the databases were fairly large. The operating system was Microsoft Windows 2000 Server. For general processing, Microsoft Office Premium 2000, including Word, Excel, and Access, was used. Functions included word processing and write-up, data calculation, chart creation, and data manipulation. The LSI-R data records from the County and the violation and conviction data from the BCA were imported into several Microsoft SQL 2000 Server tables using the import function on SQL. Data manipulations were done using a custom program written in Visual Basic 6. After the data were manipulated in the SQL table, the data from the SQL table were then imported into SPSS 13 Graduate Student Version for Windows. The select records function of SPSS was used to select various subsets of the population for further analysis.

Research Design

Validating the LSI-R. This was a retrospective study, performing a secondary analysis of data that had been collected by others (Bachman & Schutt, 2003). The calculations used were based on the methodology from two previous studies. Most calculations followed the methods used by Andrews in the original LSI study. The steps that were added included the calculation of the coefficient alphas for each sub-scale (Schlager & Simourd, 2007), and the creation of a regression model to look that used client age, gender, race, and LSI-R score as independent variables and violation by 1 year as the dependent variable (Flores, Lowenkamp, Holsinger, & Latessa, 2006).

The validation process consisted of calculating and comparing mean LSI-R scores for offenders who did or did not violate probation, violation rates for each risk level, coefficient alphas for each sub-scale, intercorrelations between the sub-scales, means sub-scale scores for offenders who did or did not violate probation, and the correlations between LSI-R sub-scale scores and violation. The results were then placed in tables or displayed with figures.

Dynamic Predictive Validity. This portion of the study used the same data as the validation study. The data had been collected in a repeated measures fashion, represented as O_{A1} O_{B1} O_{A2} O_{B2} etc. where O_{An} = Observations with the LSI-R and O_{Bn} = Observations by the Police and Probation Department. Intervals between LSI-R assessments were variable, from days to years, with a mode of 6 months. Arrests and probation violations between assessments were not used. The study analyses were run three times. The first analyses using Sample 2 could be represented as O_{A1} O_{A2} O_{B2}, the second analyses with Sample 3 as O_{A1} O_{A2} O_{A3} O_{B3}, and the third analyses with Sample 4 as O_{A1} O_{A2} O_{A3} O_{A4} O_{B4}. No longitudinal analyses were done and all tests were simply done to measure the differences in accuracy between the LSI-R assessments. For instance, with Sample 4 the scores from LSI-R #1, LSI-R #2, LSI-R #3, and LSI-R #4 were all analyzed to determine the AUC values and correlation rates for each when predicting probation violation in the year after the fourth assessment. The results for each LSI-R assessment were compared with each test to determine which set of LSI-R assessment scores were the best predictors of parole violation.

The analyses performed in this study were modeled after several previous studies. 1.) The display of the change scores was done simply to show the relative frequency of each size score change. 2.) The display of mean final score by initial score was done to determine whether a regression toward the mean was occurring. This was suggested by comments made by Raynor who had indicated that regression to the mean might occur between assessments. 3.) The cross tabulation of violation rates by initial and follow-up LSI-R risk level was a replication of the method used by Andrews and Robinson (1984). 4.) The AUC values and the correlation rates for each assessment were calculated in an attempt to measure the change in predictive validity between assessments. The logic behind this comparison was that if the new scores are better predictors than the old scores, the AUC values and correlation rates should be higher for the subsequent LSI-R assessments. 5.) The next analysis performed was a replication of the methods used by Raynor. In order to avoid regression to the mean, Raynor split the scores into two portions, above and below average, and looked at the recidivism rates for offenders with increasing and decreasing scores for each portion. He also looked at all offenders with changing scores and compared recidivism rates based upon whether scores were increasing or decreasing. Violation rates for offenders whose scores did not change were displayed for the sake of completeness. 6.) A regression model was calculated using client age, gender, race, previous LSI-R score and the change in LSI-R scores as independent variables and probation violation by 1 year after assessment as the dependent variable. Lowenkamp (personal communication, April, 2007) suggested the format for the logistic regression model.

The process described above was repeated two more times using Samples 3, and 4. The logic behind repeating the same process was that if scores changed between LSI-R #1 and LSI-R #2, then a similar process should occur between the other assessments. An additional step was added to some of the subsequent tests to compare the differences in predictive validity between all of the LSI-R assessments.

Procedure

Custom data fields and calculations. The initial database contained the LSI-R records. Additional fields were added to the LSI-R SQL table as needed and populated with data from the BCA, court services, or computations from other fields. For example, a set of fields called "A1" and "A2" were created and populated with a 1 or a 0. A 1 in "A1" represented an arrest resulting in conviction in the first 6 months. A 1 in "AV2" represented an arrest within 1 year. A set of codes for probation violations was also created using the court services data. A race field was added called "racecode." The "racecode" field was populated with a 1 if the race variable in the BCA data was coded as a "W" and a 0 for all other BCA codes. This process of culling data from the BCA tables was repeated for all variables of interest.

Some variables in the SQL table were calculated from other variables. A set of fields called "AV1" and "AV2" was created and a 1 or a 0 was placed in that field if any probation violation occurred by the end of the 6 months or 1 year respectively. A "scorecategory" variable was created and was assigned a number from 1 through 5 which represented the LSI-R risk levels, 0-11, 12-18, 19,24, 25-31, and 32-54.

Statistical analysis. The data was imported from SQL into SPSS in order to perform the statistical analyses. For sample selection, the custom fields generated by Visual Basic were used in the select cases mode of SPSS. The output results were either transcribed manually into this report or copied into a Microsoft Excel file for further manipulation. Microsoft Excel was used to create the graphs in the results.

Calculating the area under the curve. A Microsoft Excel spreadsheet was used to create the AUC values, and 95% confidence intervals (CI) in the results. The spreadsheet was originally developed by Watkins (2000) and was adapted for use with the LSI-R violation frequency scores. The magnitude of the AUC value, which ranges from 0 to 1, indicates the accuracy of the test. An AUC value of .50 would indicate a chance probability of a correct prediction. The probability of two curves being different was calculated with an Excel spreadsheet using the method suggested by Hanley and McNeil (1983) for comparing two ROC curves derived from the same cases.

The t-test for correlation coefficients. The t-test used for determining whether the correlation coefficients were different from each other was written into an excel spreadsheet using the formula provided by Blalock (1972). The formula used was

$$t = (r_{xy} - r_{zy})^{*}\ \mathrm{SQRT}[\{(n - 3)(1 + r_{xz})\}/\ \{2(1 - r_{xy}^{2} - r_{xz}^{2} - r_{zy}^{2} + 2r_{xy}{*}r_{xz}{*}r_{zy})\}],$$

where r_{xy} and r_{zy} are the correlation coefficients between the LSI-R scores and probation violation and r_{xz} is the correlation coefficient between the two LSI-R scores. The significance level was read from a t-table.

Chapter IV

RESULTS AND DISCUSSION

VALIDATING THE LSI-R

Sample 1: Mean LSI-R Scores and Probation Violation Statistics

Sample 1 was divided by probation violation status at 6 months and 1 year. The numbers per group, means, standard deviations, and median LSI-R scores for the groups are shown in Table 5. The offenders who violated probation had significantly higher LSI-R scores than offenders who didn't violate probation ($p<.001$).

Table 5

Ns, Means, Standard Deviations, and Median Values of the LSI-R #1 Scores by Probation Violation Status at Six Months and One Year for Sample 1 (N=3,190)

			LSI-R Scores		
	N	Mean		SD	Median
Six Months					
New Violation	593	26.7		9.3	27
No Violation	2597	20.9	p<.001[a]	9.6	21
One-Year					
New Violation	890	26.0		9.4	26
No Violation	2300	20.0	p<.001[a]	9.5	20
Total Offenders	3190	21.9		9.8	22

Note: SD – Standard deviation, [a] t-test probability of no difference in mean

Sample 1: AUC Values and Correlation Rates between LSI-R #1 and Probation Violation

The AUC values along with the 95% confidence intervals for the AUC, correlation rates, and the probability of no correlation between the LSI-R #1 scores and violation status were calculated for the 6 month and 1 year violation status and placed in Table 6. While the Pearson correlation between LSI-R scores and violation at 1 year ($r=.265$, $p<.001$) was higher than the 6 month rate ($r=.227$, $p<.001$), the AUC value of 67.11% for the 1 year total was found to be almost identical to the 6 month value of 66.99%. This result suggests that the predictive power of the LSI-R was stable over time. The lower bounds of the AUC values of over 64% indicate that the LSI-R was able to predict probation violation at better than chance levels.

Table 6

AUC Values, Correlation Rates, and Probability of No Correlation between LSI-R #1 and Probation Violation Status at Six Months and One Year for Sample 1 (N=3,190)

		95% C.I.[a]			
Interval	**AUC**	**Lower**	**Upper**	**r**	**p**
Six Months	66.99	64.43	69.55	.227	.000
One-Year	67.11	64.94	69.29	.265	.000

Note: [a] 95% confidence interval for AUC values

Sample 1: Rate of Probation Violation by
LSI-R Risk Level

The rates of violation at 6 months and 1 year after assessment were calculated for each risk level and placed in Figure 2. The violation rates appear to be as expected. The violation rates show a fairly linear trend from lowest risk level to highest. The 1 year probation violation rates appear to be substantially larger than the 6 month rates.

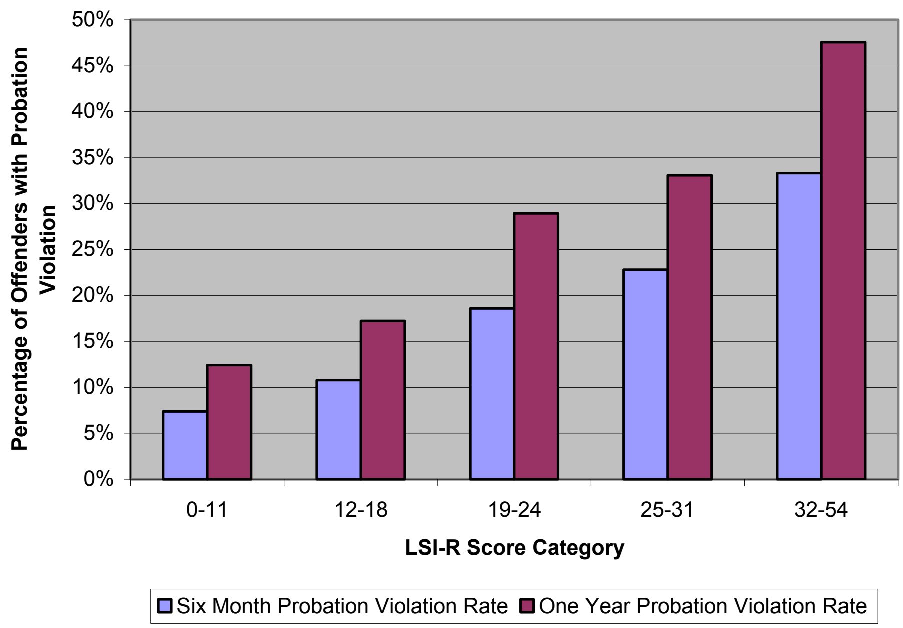

Figure 2

Six-Month and One-Year Probation Violation Rates by LSI-R Score Category for Offenders in Sample 1 (N=3,190)

The numbers of offenders in each score category, the probation violation rates at 6 months and 1 year, and the numbers of offenders who violated probation in each score category at listed in Table 7. The rate of violation increases for each risk level, and each level increases by a similar amount. The violation rates at 1 year are approximately 40% higher than at 6 months.

Table 7

Six-Month and One-Year Probation Violation Rates for Offenders in Sample 1 by LSI-R Risk Level (N=3,190)

	LSI-R Risk Level					
	0-11	**12-18**	**19-24**	**25-31**	**32-54**	**Total**
N	515	713	715	671	576	3190
Six Months						
Violation Rate	7%	11%	19%	23%	33%	19%
Probation Violations	38	77	133	153	192	593
One Year						
Violation Rate	12%	17%	29%	33%	48%	28%
Probation Violations	64	123	207	222	274	890

Sample 1: Logistic Regression Model
Predicting Probation Violation

The logistic regression model was calculated using age, gender, race, and LSI-R scores from Sample 1 and placed in Table 8. Gender and race were both coded as categorical values. Gender was coded as 1 for male and 0 for female. Race was coded as a 1 for white and 0 for non-white. While all values where significant, the *Exp(B)* values only exceeded 1 for race and the LSI-R scores. Gender appeared to produce the weakest contribution to the model.

Table 8

Logistic Regression Model Predicting Probation Violation for Sample 1 (N=3,190)

							95% C.I. for Exp(B)	
	B	**S.E.**	**Wald**	**df**	**Sig.**	**Exp(*B*)**	**Lower**	**Upper**
Age	-.026	.004	38.729	1	.000	.975	.967	.983
Gender	-.365	.112	10.607	1	.001	.694	.557	.865
Race	.590	.101	34.109	1	.000	1.805	1.480	2.200
LSI-R #1	.058	.004	167.151	1	.000	1.059	1.050	1.069
Constant	-1.520	.172	78.022	1	.000	.219		

Note: -2 log likelihood = 3465.790, $X^2(4) = 311.192$; p<.001, Cox & Snell $R^2 = .093$, Nagelkerke $R^2 = .134$

Sample 1: Inter-correlations between
LSI-R Sub-scales

The correlation coefficients for the relationships between the 10 LSI-R sub-scales were calculated and placed in Table 9. All items were significantly correlated, although none seemed to be excessively high. The coefficient alphas were also calculated for each sub-scale. The coefficient alphas seemed to be in the acceptable ranges.

Table 9

Coefficient Alphas, Inter-correlations, Means, and Standard Deviations for the LSI-R Sub-scale Scores and the LSI-R Total Scores for Sample 1 (N=3,190)

LSI-R Sub-scale	1	2	3	4	5	6	7	8	9	10	Tot.
1. Criminal History	(.77)										
2. Education/Empl.	.36	(.84)									
3. Financial	.23	.49	(.41)								
4. Family/Marital	.26	.32	.31	(.49)							
5. Accommodation	.26	.38	.30	.34	(.59)						
6. Leisure/Recreation	.21	.39	.31	.24	.27	(.56)					
7. Companions	.36	.36	.17	.25	.31	.23	(.63)				
8. Alcohol/Drugs	.33	.29	.24	.17	.25	.22	.29	(.86)			
9. Emotional/Personal	.20	.23	.32	.35	.18	.12	.16	.21	(.64)		
10. Attitude/Orientation	.40	.37	.25	.28	.31	.37	.32	.35	.16	(.78)	
LSI-R Total	.67	.75	.54	.53	.54	.49	.57	.65	.46	.63	(.90)
Mean	4.3	3.8	1.1	1.7	0.7	1.2	1.8	4.0	2.1	1.3	21.9
Standard Deviation	2.3	3.0	0.7	1.2	0.9	0.8	1.2	2.8	1.4	1.4	9.8

Note: Coefficient alphas are in parenthesis. All correlations are significant at the .001 level.
LSI-R = Level of Service Inventory – Revised.

Sample 1: Means and Correlation between
LSI-R Sub-scale Scores and
Probation Violation

The mean LSI-R sub-scale scores with standard deviation, and correlations between LSI-R sub-scale and 1 year violation rate were calculated and displayed in Table 10. All sub-scales appear to be correlated with recidivism at significant levels. The highest correlation with violation is for criminal history (r=.236) and the lowest is for the Emotional/Personal sub-scale (r=.036), which is the only sub-scale that is not significant at the p<.001 level (p=.043).

Table 10

Mean LSI-R Sub-scale Scores and the Correlation with Probation Violation within One Year after Assessment for Offenders in Sample 1 (N=3,190)

	Violation		**No Violation**			
	Mean	**SD**	**Mean**	**SD**	**r**	**p**
LSI-R Sub-Level						
Criminal History	5.2	2.2	4.0	2.3	.236	.000
Education/Employment	4.8	2.9	3.4	2.9	.214	.000
Financial	1.2	0.7	1.0	0.8	.100	.000
Family/Marital	1.9	1.2	1.6	1.2	.127	.000
Accommodation	1.0	1.1	0.6	0.9	.176	.000
Leisure/Recreation	1.4	0.8	1.2	0.8	.134	.000
Companions	2.1	1.3	1.6	1.2	.168	.000
Alcohol/Drugs	4.6	2.8	3.8	2.8	.117	.000
Emotional/Personal	2.2	1.4	2.1	1.4	.036	.043
Attitude/Orientation	1.8	1.5	1.1	1.4	.226	.000
LSI-R Total	26.1	9.4	20.3	9.5	.265	.000

Note: SD – Standard Deviation

Sample 2: Changes in Scores between
LSI-R #1 and LSI-R #2

To determine how dynamic changes in LSI-R scores are related to subsequent violation rates, the changes in LSI-R scores between LSI-R #1 and LSI-R #2 were calculated for Sample 2 and graphed in Figure 3. Changes in LSI-R scores ranged from -21 to 27 with a mean change of –1.43 (Std. Dev. = 6.8), the median was -1, and the mode was 0. The mean number of days between assessments was 257 with a significant amount of variation (StdDev= 129). The median number of days between assessments was 222, and the mode was 181, scores ranged from 0 to 996 days.

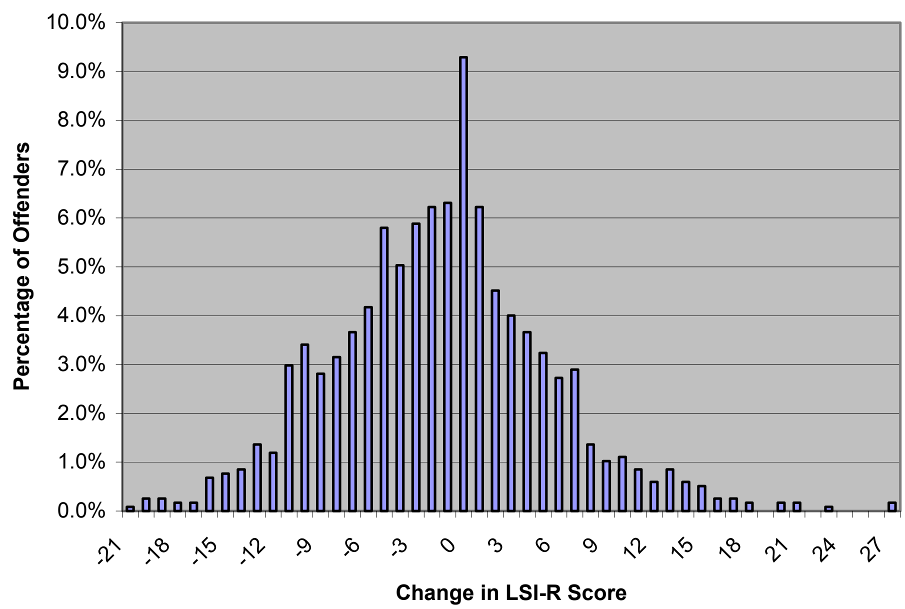

Figure 3

Percentage of Offenders by Change in LSI-R Score for Offenders in Sample 2
(N=1,173)

Sample 2: Regression to the Mean

Raynor had suggested that there might be a regression toward the mean in LSI-R score changes between LSI-R #1 and LSI-R #2. To test that hypothesis, the mean scores of the LSI-R #2 assessments were plotted as a function of the LSI-R #1 scores and the distribution was plotted in Figure 4.

There was a marked tendency for below average scores on LSI-R #1 to be higher on the LSI-R #2 assessment and above average scores on LSI-R #1 to be slightly lower on the LSI-R #2 assessment. This suggests that there is a regression towards the mean in the LSI-R scores between LSI-R #1 and LSI-R #2.

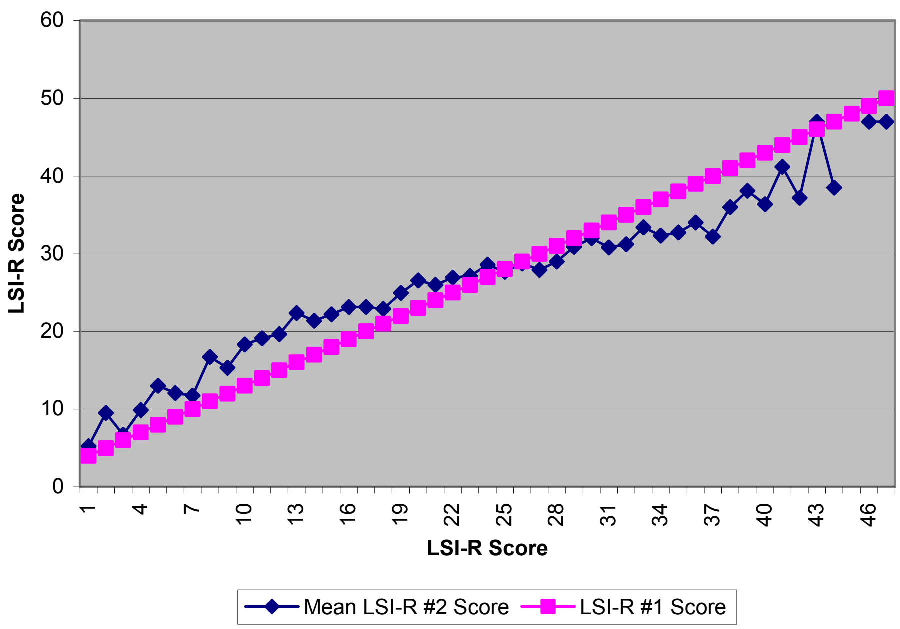

Figure 4

Mean LSI-R Score at LSI-R #2 Plotted for Each LSI-R #1 Score for Offenders in Sample 2 (N=1,173)

<u>Sample 2: Prediction Accuracy for
LSI-R#1 and LSI-R #2</u>

In the study done by Andrews and Robinson, the score distributions for the first and second assessments were broken down by risk level and compared to see which scores were the more accurate predictors of outcome. The numbers and percentages of violations at 1 year were plotted for both LSI-R #1 and LSI-R #2 at each risk level (1-11, 12-18, 19-24, 25-31, and 32-54) for Sample 2 and placed in Table 11. There were overall improvements in scoring at the 0-11 level (5% vs. 10%), at the 25-31 level (31% vs. 24%), and the highest level (41% vs. 39%). The percentages tend to increase from right to left as expected. This suggests that LSI-R #2 was a better predictor of risk than LSI-R #1 due to more accurate assessment of risk level.

Table 11

Probation Violation Rates after LSI-R #2 for LSI-R #1 and LSI-R #2

LSI-R #1 Risk Level	**LSI-R #2 Risk Level**					
	0-11	**12-18**	**19-24**	**25-31**	**32-54**	**Overall**
0-11	4% (2/47)	16% (3/19)	50% (4/8)	0% (0/1)	0% (0/2)	12% (9/77)
12-18	9% (2/23)	9% (7/76)	33% (12/36)	20% (3/15)	100% (3/3)	18% (27/153)
19-24	0% (0/9)	17% (15/90)	17%(21/127)	29% (22/76)	45% (10/22)	21% (68/324)
25-31	0% (0/5)	16% (8/51)	18% (20/113)	35%(44/125)	49% (27/55)	28% (99/349)
32-54		17% (1/6)	34% (15/44)	45% (37/83)	53% (72/137)	46%(125/270)
Overall	5% (4/84)	14% (34/242)	22% (72/328)	35% (106/300)	51% (112/219)	28%(328/1173)

Note: Numbers in parenthesis – (Probation violation count/Total offenders)

Sample 2: Comparison of AUC Values

In order to get a better understanding of the differences in predictive value for the two LSI-R assessments, the probation violation rates at 1 year after LSI-R #2 were calculated by LSI-R score level for both LSI-R #1 and LSI-R #2 and placed in Table 12 below, along with the AUC values, 95% confidence intervals, standard errors, and probability of difference between the AUC values.

The results appear to be as expected. The score distribution for LSI-R #2 looks better than for LSI-R #1. The AUC value for LSI-R #2 is significantly larger than the AUC value for LSI-R #1 (71.01 vs. 65.74) indicating that LSI-R #2 is a significantly better predictor of probation violation than LSI-R #1.

Table 12

Probation Violation Rates by LSI-R Risk Level, AUC Values, 95% Confidence Limits, and the Probability of No Difference in AUC Values for LSI-R Scores and Probation Violation Rates by One Year after LSI-R #2 (N=1,173)

	LSI-R Score Level							**95% C.I.**[a]			
LSI-R #	**0-11 %**	**12-28 %**	**19-24 %**	**25-31 %**	**32-54 %**	**Total %**	**AUC**	**Lower**	**Upper**	**S.E.**[b]	**p**[c]
LSI#1	12	18	21	28	46	28	65.74	62.12	69.35	.0178	
LSI#2	5	14	22	35	51	28	71.01	67.52	74.49	.0185	.001

a: 95% Confidence limits for AUC value. b: Standard error of AUC; c: Probability of no difference between the AUC value in that row and the AUC value in the row above.

Sample 2: Comparison of Correlation Rates

The correlation rates between LSI-R #1 and LSI-R #2 and probation violation were calculated, along with the correlation between the LSI-R #1 and LSI-R #2 scores and placed in Table 13 below. A t-test was done to test whether the two correlation rates were significantly different from each other. The results were exactly as expected. The correlation rate between LSI-R scores and probation violation was significantly higher for LSI-R #2 than LSI-R #1 (.338 vs. .243). The correlation between the two sets of LSI-R scores was moderate.

Table 13

Correlation Rates between LSI-R Scores and Probation Violation Rates by One Year after LSI-R #2, the Inter-correlation between LSI-R Assessments, and the t-test Values Indicating the Probability of No Difference between Correlation Rates for Sample 2 (N=1,173)

	Probation Violation	**LSI-R #1**	**LSI-R #2**
Probation Violation			
LSI-R #1	.236[a]		
LSI-R #2	.329	.684 (t_{1-2}=3.03, p<.005)[b]	

a. All correlations are significant at p<.001 b. t-test value for difference in correlation rates and probability of no difference.

Sample 2: The Effect of Changes in LSI-R Scores from LSI-R #1 to LSI-R #2

This test replicates the methods used by Raynor (2007). The probation violation rates were calculated for offenders with increasing and decreasing scores for both above average and below average LSI-R #1 scores and placed in Table 14 below. For the sake of completeness, scores that did not change were included. A comparison was also made between all offenders with increasing scores and all offenders with increasing scores. The results are as expected. Increasing scores at each initial level were associated with significantly higher levels of violation when compared with decreasing scores. The overall set of offenders with increasing scores had significantly higher violation rates than offenders with decreasing scores.

Table 14

Probation Violation Rates and Mean Values for LSI-R #1 and LSI-R #2 for Increasing, Decreasing, Static, and Total Sample for Sample 2

LSI-R Change Category	N	1 Yr Violation Rate	Mean LSI-R #1	Mean LSI-R #2
LSI-R #1 <=25 Increasing	277	26%	18.2	24.2
LSI-R #1 <=25 Decreasing	270	13% $p<.001^a$	20.4 $p<.001^b$	15.6 $p<.001^b$
LSI-R #1 > 25 Increasing	139	54%	32.1	36.7
LSI-R #1 > 25 Decreasing	378	32% $p<.001^a$	32.5 ns^b	25.6 $p<.001^b$
LSI-R #1 = LSI-R #2	109	24%	21.8	21.8
Total	1173	28%	25.3	23.9
All Increasing	416	35%	22.9	28.4
All Decreasing	648	24% $p<.001^a$	27.5 $p<.001^b$	21.3 $p<.001^b$

Note: [a] Mann Whitney U probability of no difference, [b] t-test probability of no difference in mean

Sample 2: Logistic Regression Model
Predicting Probation Violation

The logistic model predicting violation for Sample 2 was calculated and placed in Table 15. In addition to age gender and race, the LSI-R #1 score was used as a base and the change in score from LSI-R #1 to LSI-R #2 was added. The model shows that the change in score was a significant factor in predicting subsequent violation.

Table 15

Logistic Regression Model Predicting Probation Violation for Sample 2 (N=1,173)

							95% C.I. for Exp(*B*)	
	B	**S.E.**	**Wald**	**df**	**Sig.**	**Exp(*B*)**	**Lower**	**Upper**
Age	-.024	.007	12.071	1	.001	.976	.963	.990
Gender	-.096	.188	.261	1	.609	.909	.629	1.312
Race	.320	.169	3.562	1	.059	1.377	.988	1.919
LSI-R #1	.088	.010	79.176	1	.000	1.092	1.071	1.114
Change	.077	.011	48.161	1	.000	1.080	1.057	1.104
Constant	-2.440	.364	44.976	1	.000	.087		

Note: -2 log likelihood = 1244.011, $X^2(5) = 146.225$; $p<.001$, Cox & Snell $R^2 = .117$, Nagelkerke $R^2 = .169$

<u>Sample 3: Changes in Scores between
LSI-R #2 and LSI-R #3</u>

Of the 1173 offenders with a second assessment in Sample 2, a group of 616 offenders, which will be referred to as Sample 3, had a third LSI-R assessment before 2005. Changes in LSI-R scores for this group of offenders were calculated for the change between LSI-R #1 and LSI-R #3, and between LSI-R #2 and LSI-R #3, and the percentage of offenders at each change level was graphed in Figure 5 below. The distribution for changes in LSI-R scores between LSI-R #2 and LSI-R #3 is peaked at the score change = 0. The mean change between LSI-R #2 and LSI-R #3 was -0.77 (Std.Dev.=5.3). The mean days between assessments two and three was 215 days (StdDev. = 90), with a mode of 179 days.

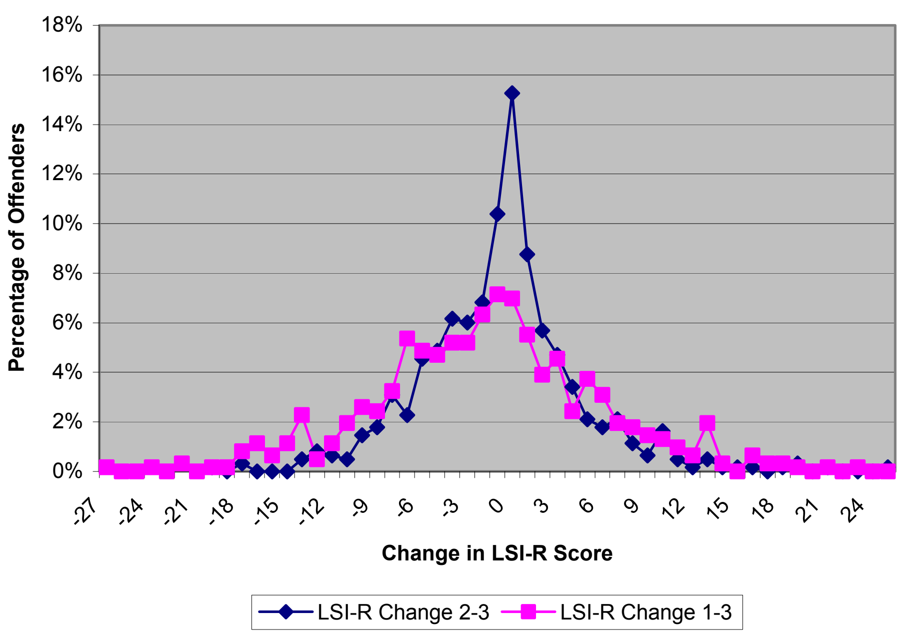

Figure 5

Percentage of Offenders by Change in LSI-R Score for Sample 3 (N=616)

<u>Sample 3: Regression to the Mean</u>

The mean scores for LSI-R #3 were plotted for each score at LSI-R #2 and plotted in Figure 6 below. Visual observation indicates that either the mean scores remained the same or there was a continued regression toward the mean, with above average scores getting lower and below average scores getting higher. The regression for above average scores seems to be larger than for below average scores.

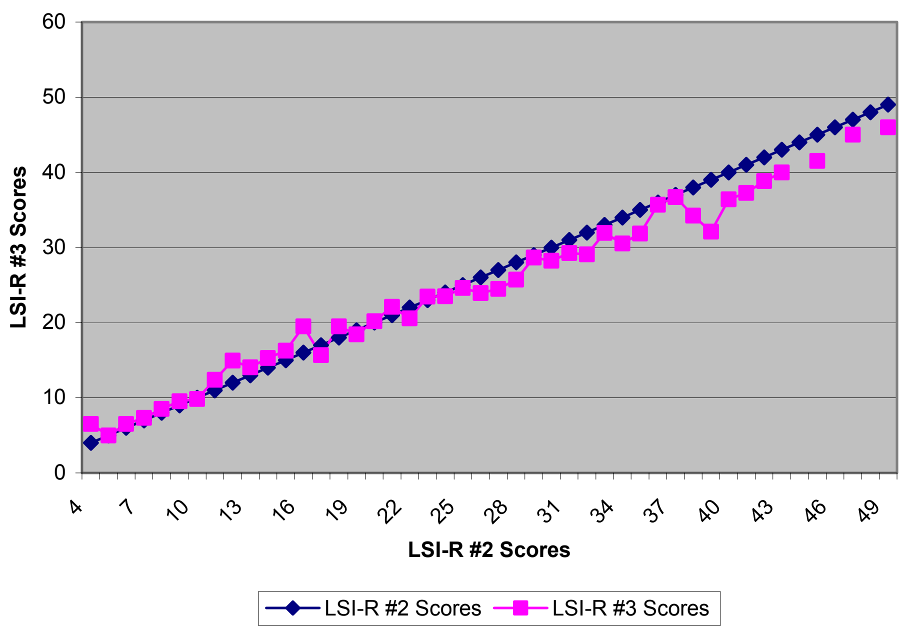

Figure 6

Mean LSI-R Score at LSI-R #3 Plotted for Each LSI-R #2 Score for Offenders in Sample 3 (N=616)

Sample 3: Prediction Accuracy for
LSI-R #2 and LSI-R #3

The numbers and percentages of violations at 1 year were plotted for both LSI-R #2 and LSI-R #3 and placed in Table 16. The overall improvements in scoring were modest. The overall results for LSI-R #3 and LSI-R #2 are very similar with slight improvements for LSI-R #3 over LSI-R #2. An examination of the detail shows that low scorers in LSI-R #2 that were assessed higher in LSI-R #3 tended to be placed in accurate categories but high scorers that were assessed at a lower risk level still tended to violate probation at high rates. This seems to indicate that, for this assessment, increases in LSI-R score are predictive of increased offending, but decreases in LSI-R score don't predict decreases in offending.

Table 16

Probation Violation Rates by One Year after LSI-R #3 for LSI-R #2 and LSI-R #3

LSI-R #2 Risk Level	**LSI-R #3 Risk Level**					
	0-11	**12-18**	**19-24**	**25-31**	**32-54**	**Overall**
0-11	3% (1/34)	13% (1/8)	-	-	-	5% (2/42)
12-18	0% (0/12)	6% (4/72)	14% (3/21)	20% (1/5)	67% (2/3)	9% (10/113)
19-24	0% (0/2)	17% (8/47)	19% (16/85)	32% (10/31)	75% (6/8)	23% (40/173)
25-31	-	7% (1/15)	30% (16/53)	26% (21/81)	46% (12/26)	29% (12/26)
32-54	-	50% (1/2)	33% (1/3)	35% (11/31)	49% (38/77)	45% (38/77)
Overall	2% (1/48)	10% (15/144)	22%(36/162)	29% (43/148)	51% (58/114)	25% (153/616)

Note: Numbers in parenthesis – (Probation violation count/Total offenders)

Sample 3: Comparison of AUC Values

The probation violation rates at 1 year after LSI-R #3 were calculated by LSI-R risk level for the first three LSI-R assessments. These results, along with the AUC values, 95% confidence intervals, standard error values, and probabilities of no difference in AUC value for each LSI-R assessment are shown in Table 17 below. The overall classification of scores appears to be improved from LSI-R #1 to LSI-R #2 and from LSI-R #2 to LSI-R #3. The AUC values indicate that scores for LSI-R #2 and LSI-R #3 are significantly ($p_{1-2} = .003$, $p_{1-3} = .001$) more predictive of violation than the scores from LSI-R #1, but there is no significant difference between the AUC values for LSI-R #2 and LSI-R #3 ($p_{2-3} = .215$). These findings indicate that the improvement in predictive validity that occurred between LSI-R #1 and LSI-R #2 was not repeated between LSI-R #2 and LSI-R #3.

Table 17

Probation Violation Rates by LSI-R Risk Level, AUC Values, 95% Confidence Limits, and the Probability of No Difference in AUC Values for LSI-R Scores and Probation Violation Rates by One Year after LSI-R #3 (N=616)

	LSI-R Risk Level							**95% C.I.**[a]			
LSI-R #	**0-11 %**	**12-28 %**	**19-24 %**	**25-31 %**	**32-54 %**	**Total %**	**AUC**	**Lower**	**Upper**	**S.E.**[b]	**p**[c]
LSI#1	7	14	22	29	36	25	64.39	59.13	69.64	.0268	
LSI#2	5	9	23	29	45	25	70.62	65.57	75.67	.0258	.003
LSI#3	2	10	22	29	51	25	72.49	67.53	77.46	.0253	.215

a: 95% Confidence limits for AUC value. b: Standard error of AUC; c: Probability of no difference between the AUC value in that row and the AUC value in the row above.

<u>Sample 3: Comparison of Correlation Rates</u>

The correlation rates between the first three LSI-R assessment scores and probation violation rates by 1 year after LSI-R #3 were calculated, along with the inter-correlation rates between the three LSI-R assessments, the t-values from a difference test, and the probability of no difference between the correlation rates and placed in Table 18 below. The results of the t-tests indicate that the correlation rates between the LSI-R #2 and LSI-R #3 scores and probation violation were significantly higher than the correlation rate for the LSI-R #1 scores ($p_{1-2} < .005$, $p_{1-3} < .005$), but they were not significantly different from each other ($p_{2-3} = .100$). This indicates that there was no significant improvement in predictive validity between LSI-R #3 and LSI-R #2.

Table 18

Correlation Rates between LSI-R Scores and Probation Violation Rates by One Year after LSI-R #3, the Inter-correlation between LSI-R Assessments, and the t-test Values Indicating the Probability of No Difference between Correlation Rates for Sample 3 (N=616)

	Probation Violation	**LSI-R #1**	**LSI-R #2**	**LSI-R #3**
Probation Violation				
LSI-R #1	.223[a]			
LSI-R #2	.317	.719 (t_{1-2}=3.29, p<.005)[b]		
LSI-R #3	.348	.635 (t_{1-3}=2.88, p<.005)	.811 (t_{2-3}=1.32, p<.100)	

a. All correlations are significant at p<.001 b. t-test value for difference in correlation rates and probability of no difference.

Sample 3: The Effect of Changes in LSI-R Scores from LSI-R #2 to LSI-R #3

To determine how changes in score level affected prediction of violation, the probation violation rates for increasing and decreasing LSI-R scores were calculated for both above and below average scores on LSI-R #2 and shown in Table 19. The numbers are not quite as expected. While increasing scores are accompanied by significantly higher probation violation rates than are decreasing scores, for initial LSI-R scores that were less than average, the difference was not significant for the above average scores. This suggests that, the improvement in predictive validity between LSI-R #2 and LSI-R #3 was not consistent for the entire score range.

Table 19

Probation Violation Rates and Mean Values for LSI-R #2 and LSI-R #3 for Increasing, Decreasing, Static, and Total Sample for Sample 3

LSI-R Change Category	N	1 Yr Violation Rate	Mean LSI-R #2	Mean LSI-R #3
LSI-R #2 <=25 Increasing	136	24%	18.9	23.7
LSI-R #2 <=25 Decreasing	154	16% p<.05[a]	19.6 ns[b]	16.0 p<.001[b]
LSI-R #2 > 25 Increasing	76	46%	32.1	35.9
LSI-R #2 > 25 Decreasing	156	31% ns[a]	32.1 ns[b]	26.7 p<.001[b]
LSI-R #2 Same as LSI-R #3	94	14%	19.4	19.4
Total	616	25%	24.14	23.37
All Increasing	212	32%	23.7	28.1
All Decreasing	310	24% p<.05[a]	25.9 p<.01[b]	21.4 p<.001[b]

Note: [a] Mann Whitney U probability of difference, [b] t-test probability of difference

Sample 3: Logistic Regression Model
Predicting Probation Violation

The logistic model predicting violation for Sample 3 was calculated and placed in Table 20. In addition to age gender and race, the LSI-R #2 score was used as a base and the change in score from LSI-R #2 to LSI-R #3 was added. The model shows that the changes in scores from LSI-R #2 to LSI-R #3 were a significant factor in predicting subsequent violation.

Table 20

Logistic Regression Model Predicting Probation Violation for Sample 3 (N=616)

							95% C.I. for Exp(*B*)	
	B	**S.E.**	**Wald**	**df**	**Sig.**	**Exp(*B*)**	**Lower**	**Upper**
Age	-.024	.010	5.297	1	.021	0.977	0.957	0.997
Gender	-.324	.272	1.411	1	.235	0.723	0.424	1.234
Race	.588	.234	6.325	1	.012	1.800	1.139	2.846
LSI-R #2	.097	.014	47.162	1	.000	1.102	1.072	1.133
Change	.067	.019	12.848	1	.000	1.069	1.031	1.108
Constant	-2.851	.544	27.415	1	.000	0.058		

Note: -2 log likelihood = 598.632, $X^2(5) = 91.959$; p<.001, Cox & Snell $R^2 = .139$, Nagelkerke $R^2 = .206$

Sample 4: Changes in Scores between
LSI-R #3 and LSI-R #4

Of the 616 offenders with a third assessment in Sample 3, a group of 285 offenders, which will hereafter be referred to as Sample 4, had a fourth assessment before 2005. Changes in LSI-R scores for this group of offenders were calculated for the change between LSI-R #1 and LSI-R #4, and between LSI-R #3 and LSI-R #4, and the percentage at each change level was graphed in Figure 7 below. The distribution for changes in LSI-R scores between LSI-R #3 and LSI-R #4 is peaked at the score change = 0. The mean change between LSI-R #3 and LSI-R #4 was -0.65 (Std.Dev.=4.9), with scores ranging from –17 to 23. The mean days between assessments three and four was 192 days (StdDev. = 77), with a mode of 187.

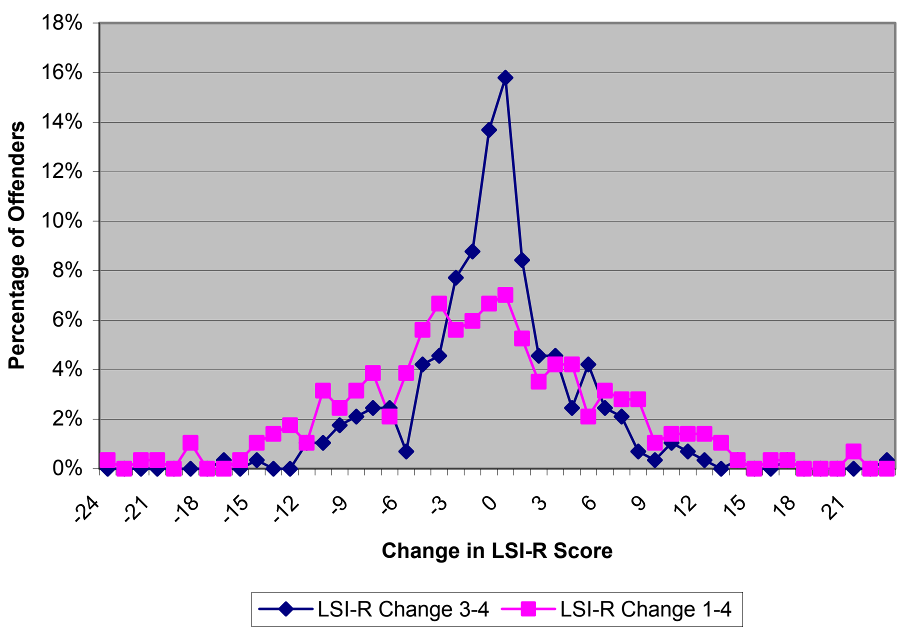

Figure 7

Percentage of Offenders by Change in LSI-R Score for Sample 4 (N=285)

Sample 4: Regression to the Mean

The mean scores for LSI-R #4 were plotted for each score at LSI-R #3 and plotted in Figure 8 below. Visual observation indicates that either the mean scores remained the same or there was a continued regression toward the mean, with above average scores generally getting lower and below average scores getting higher. The regression to the mean for above average scores seems to be larger than for below average scores.

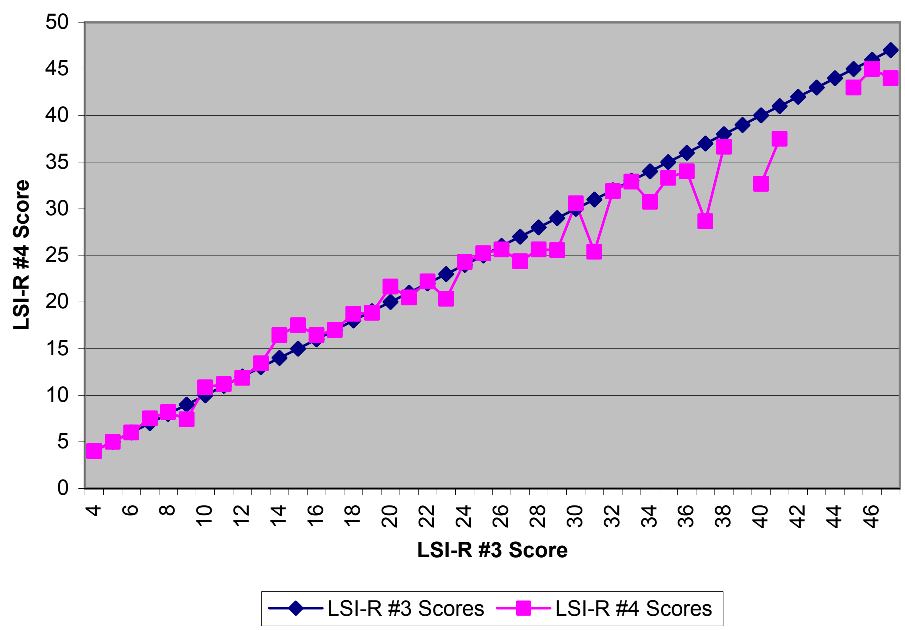

Figure 8

Mean LSI-R Score at LSI-R #4 Plotted for Each LSI-R #3 Score for Offenders in Sample 4 (N=285)

Sample 4: Prediction Accuracy for
LSI-R #3 and LSI-R #4

The numbers and percentages of violations at 1 year were plotted for both the LSI-R #3 and LSI-R #4 assessments at each risk level for Sample 4 and placed in Table 21. It is difficult to tell if there are overall improvements in scoring. The overall results for LSI-R #4 and LSI-R #3 are very similar with a slight improvement for LSI-R #4 over LSI-R #3 in the 25-31 score level and a slight decline in accuracy for LSI-R #4 over LSI-R #3 in the 32-54 score level. An examination of the detail shows that there appear to be problems with both increasing and decreasing scores that do not seem to match the expected violation rate for that level. The probation violation rates should increase from left to right for all risk levels and that is not what happened.

Table 21

Probation Violation Rates by One Year after LSI-R #4 for LSI-R #3 and LSI-R #4

LSI-R #3 Risk Level	**LSI-R #4 Risk Level**					
	0-11	**12-18**	**19-24**	**25-31**	**32-54**	**Overall**
0-11	4% (1/24)	0% (0/3)	-	-	-	4% (1/27)
12-18	0% (0/6)	15% (6/41)	22% (2/9)	67% (2/3)	100% (1/1)	18% (11/60)
19-24	0% (0/1)	25% (4/16)	17% (6/35)	56% (10/18)	0% (0/3)	27% (20/73)
25-31	-	14% (1/7)	24% (5/21)	36% (15/42)	25% (2/8)	29% (2/8)
32-54	-	-	100% (1/1)	50% (9/18)	64% (18/28)	60% (28/47)
Overall	3% (1/31)	16% (11/67)	21% (14/66)	44% (36/81)	53% (21/40)	29% (83/285)

Note: Numbers in parenthesis – (Probation violation count/Total offenders)

Sample 4: Comparison of AUC Values

The probation violation rates by 1 year after LSI-R #4 were calculated by LSI-R risk level for each LSI-R assessment and placed in Table 22, along with the AUC values, 95% confidence intervals, standard errors, and probabilities of no difference in AUC values for each LSI-R assessment. The last three sets of LSI-R assessment scores are significantly better at predicting risk than LSI-R #1 but it is difficult to determine visually which of the last three sets of risk scores is the better predictor. The AUC values for the last three LSI-R scores are significantly higher than the AUC value for LSI-R #1 (p_{1-3} = .002, p_{1-3} = .002, p_{1-4} = .003), but there is no significant difference between any of the other AUC values (p_{2-3} = .492, p_{2-4} = .271, p_{3-4} = .264). This indicates that the gain in predictive validity that occurred between LSI-R #1 and LSI-R #2 was not repeated between any of the other LSI-R assessments.

Table 22

Probation Violation Rates by LSI-R Risk Level, AUC Values, 95% Confidence Limits, and the Probability of No Difference in AUC Values for LSI-R Scores and Probation Violation Rates by One Year after LSI-R #4 (N=285)

	LSI-R Risk Level							95% C.I.[a]			
LSI-R #	**0-11 %**	**12-28 %**	**19-24 %**	**25-31 %**	**32-54 %**	**Tot. %**	**AUC**	**Lower**	**Upper**	**S.E.[b]**	**p[c]**
LSI #1	14	14	36	30	40	29	62.70	55.35	70.04	.0375	
LSI #2	4	16	25	35	53	29	69.72	62.68	76.77	.0359	.008
LSI #3	4	18	27	29	60	29	69.76	62.71	76.80	.0359	.492
LSI #4	3	16	21	44	53	29	71.25	64.30	78.20	.0355	.264

a: 95% Confidence limits for AUC value. b: Standard error of AUC; c: Probability of no difference between the AUC value in that row and the AUC value in the row above.

Sample 4: Comparison of Correlation Rates

The correlation rates for the four LSI-R scores, the correlations between LSI-R scores, t-test values and probabilities of no difference between correlation rates were calculated using the probation violations by 1 year after LSI-R #4 and placed in Table 23 below. The correlation rates for LSI-R #2, LSI-R #3, and LSI-R #4 were all significantly higher than LSI-R #1, but there was no significant difference between the correlation rates for the last three sets of LSI-R scores. The inter-correlation rates between the last three score distributions were in the high range. The results indicate that the improvements in predictive validity that occurred between LSI-R #2 and LSI-R #1 were not repeated between any of the subsequent assessments.

Table 23

Correlation Rates between LSI-R Scores and Probation Violation Rates by One Year after LSI-R #4, the Inter-correlation between LSI-R Assessments, and the t-test Values Indicating the Probability of No Difference between Correlation Rates for Sample 4 (N=285)

	Probation Violation	**LSI-R #1**	**LSI-R #2**	**LSI-R #3**	**LSI-R #4**
Probation Violation					
LSI-R #1	.203[a]				
LSI-R #2	.326	.749 (t_{1-2}=2.79, p<.005)[b]			
LSI-R #3	.334	.650 (t_{1-3}=3.08, p<.005)	.841 (t_{2-3}=0.27, p<.40)		
LSI-R #4	.341	.646 (t_{1-4}=2.93, p<.005)	.802 (t_{2-4}=0.43, p<.40)	.836 (t_{3-4}=0.21, ns)	

a. All correlations are significant at p<.001 b. t-test value for difference in correlation rates and probability of no difference.

Sample 4: The Effect of Changes in LSI-R Scores from LSI-R #3 to LSI-R #4

To determine how changes in score level affected prediction of violation, the method used by Raynor was applied to the score changes between the third and fourth assessments. The results are shown in Table 24.

The numbers are not as expected. Increasing scores are accompanied by increasing violation rates when compared to decreasing scores for the lower half of the distribution but the opposite is true for the high end where increasing scores actually have a lower violation rate than the decreasing scores. The overall difference between increasing scores and decreasing scores is not significant, and is less than that seen for the changes from LSI-R #1 to LSI-R #2, and the changes from LSI-R #2 to LSI-R #3.

Table 24

Probation Violation Rates and Mean Values for LSI-R #3 and LSI-R #4 for Increasing, Decreasing, Static, and Total Sample for Sample 4

LSI-R Change Category	N	1 Yr Violation Rate	Mean LSI-R #3	Mean LSI-R #4
LSI-R #3 <=25 Increasing	62	31%	18.7	23.1
LSI-R #3 <=25 Decreasing	78	18% ns[a]	18.7 ns[b]	15.7 $p<.001$[b]
LSI-R #3 > 25 Increasing	32	41%	31.3	35.2
LSI-R #3 > 25 Decreasing	68	43% ns[a]	31.6 ns[b]	26.4 $p<.001$[b]
LSI-R #3 Same as LSI-R #4	45	18%	18.1	18.1
Total	285	29%	23.1	22.4
All Increasing	94	34%	22.9	27.2
All Decreasing	146	29% ns[a]	24.7 ns[b]	20.7 $p<.001$[b]

Note: [a] Mann Whitney U probability of difference, [b] t-test probability of difference

Sample 4: Logistic Regression Model
Predicting Probation Violation

The logistic model predicting violation for Sample 4 was calculated and placed in Table 25. In addition to age gender and race, the LSI-R #3 score was used as a base and the change in score from LSI-R #3 to LSI-R #4 was added. In this model, the change in score between assessments from LSI-R #3 to LSI-R #4 was not a significant factor in predicting subsequent violation.

Table 25

Logistic Regression Model Predicting Probation Violation for Sample 4 (N=285)

							95% C.I. for Exp(*B*)	
	B	**S.E.**	**Wald**	**df**	**Sig.**	**Exp(*B*)**	**Lower**	**Upper**
Age	-.011	.014	.590	1	.442	.989	.962	1.017
Gender	-.507	.406	1.561	1	.212	.603	.272	1.334
Race	.813	.340	5.726	1	.017	2.255	1.158	4.389
LSI-R #3	.101	.020	24.599	1	.000	1.106	1.063	1.151
Change	.046	.029	2.574	1	.109	1.047	.990	1.107
Constant	-3.049	.763	15.969	1	.000	.047		

Note: -2 log likelihood = 299.191, $X^2(5) = 44.660$; p<.001, Cox & Snell $R^2 = .145$, Nagelkerke $R^2 =$ .207

Chapter V

SUMMARY AND CONCLUSIONS

VALIDATING THE LSI-R

Summary of Findings

Although the accuracy of the LSI-R in predicting offender risk levels has been verified in many locations, it is recommended that the predictive validity of the LSI-R be verified with each group of offenders with which it is used. This study looked at the predictive validity of the LSI-R for offenders in a Midwestern community.

The LSI-R assessment data used in this study was collected from 2002 through 2004 by probation officers in a Community Corrections department as a normal part of their operations. Offenders were given initial LSI-R assessments at intake and additional LSI-R assessments whenever their situation changed. Most offenders were given additional assessments at 6 month intervals. The LSI-R data provided by the community corrections department was matched with records of arrest resulting in conviction from the Minnesota BCA and records of probation violation resulting in prison commitment from the State Court Services databases. The first LSI-R assessment records of offenders who had an assessment before 2005, along with arrest and probation violation data, were used to validate the LSI-R with this population.

This study used many of the same basic methods as the original LSI validation study done by Andrews in 1982. The LSI-R scores of the offenders were compared with probation violation rates to determine whether the score levels were related to probation violation levels. More modern methods were also used, such as the determination of the coefficient alphas, the AUC values, and calculation of logistic regression models.

A comparison of the mean LSI-R scores for the offenders who violated probation after the completion of the LSI-R with those who didn’t showed that the means for offenders who violated probation were significantly higher than the means for offenders who did not violate probation. This was true for violations made by either 6 months or 1 year after assessment. The correlation rate between the LSI-R scores and the violation rate at 1 year ($r=.265$, $p<.001$) was higher than at 6 months ($r=.227$, $p<.001$). The AUC of 67.11% for the 1 year totals was found to be almost identical to the 6 month value of 66.99%, suggesting that the predictive accuracy of the LSI-R was stable over time.

When the violation rates were compared by risk level, lower risk offenders had lower probation violation rates and higher risk offenders had higher violation rates. There was a fairly linear relationship between the probation violation rate and risk level.

A logistic regression analysis holding age, gender, and race constant showed that the LSI-R was the most significant predictor of probation violation. Race appeared to be a fairly large contributor to the logistic regression model.

The LSI-R sub-scale scores were significantly correlated with violation rates. The highest correlation with violation was for Criminal History ($r=.236$, $p<.001$) and the lowest correlation was for the Emotional/Personal sub-scale ($r=.036$, $p<.05$).

Conclusions

The veracity of the first hypothesis, "The LSI-R is a valid risk predictor for offenders served by a Midwestern County Community Corrections Department," appears to have been demonstrated by this study. The AUC values of approximately 67% show that the LSI-R scores predict risk at better than chance levels. The lower 95% confidence intervals of 64.43 and 64.94 for the AUC scores give a clear indication that this result does not appear to be due to chance.

The results indicate that the LSI-R is operating at acceptable levels for this type of offender population. The correlation rates between the LSI-R scores and violation found in this study, .227 at 6 months and .265 at 1 year, are comparable to, or slightly better than, the results found by Flores, Lowenkamp, Holsinger, and Latessa (2006) in other Midwestern community corrections settings.

Study Limitations

This study was somewhat limited by the inability to match 100% of the offenders who had LSI-R assessments with the BCA violation data. A check with the county showed that many of the offenders who weren't matched were either women who had changed last names or were misdemeanor offenders. The effect on outcome would appear to have been an increase in low risk offenders with no further arrests.

Another limitation of this study was due to an issue that is related to the arrest data used in this study. Not all offenders who commit crimes get caught. The conviction rate for crimes committed in the County used in this study was 50% in 2005 (Minnesota BCA, 2006). It is certainly probable that some offenders in this study did not make it into the probation violation status because they didn't get caught.

Recommendations for Future Research

Future research with this offender population could look at gender and racial differences in the predictive validity of the LSI-R. Some researchers have indicated that additional research is needed with female offenders who are assessed by the LSI-R to determine whether need factors differe between male and female offenders (Reisig, Holtfreter, & Morash, 2006; Holtfreter, Reisig, & Morash, 2004). Recent research by Schlager and Simourd (2007) has indicated that the predictive validity of the LSI-R may not be as high when used with Black and Hispanic populations of offenders. Scores for other race offenders in this sample could be compared with white offenders to determine whether differences between races were present.

Staff training practices could also be examined to see whether there might be room to improve the overall predictive validity of the LSI-R with this population. A more in-depth analysis might look at how prediction accuracy varies by probation officer. The results found by Flores, Lowenkamp, Holsinger, and Latessa (2006) suggest that both the training and experience of the raters affect the overall level of predictive accuracy for the LSI-R.

DYNAMIC PREDICTIVE VALIDITY OF THE LSI-R

Summary of Findings

Few studies have investigated the dynamic predictive validity of the LSI-R. Previous results were supportive of the premise that changes in LSI-R scores are related to changes in recidivism rates. There were questions about the validity and generalizability of some of the previous studies due to non-random sampling, small sample sizes, and sampling from only one country. Since these studies only measured two assessments, there was a need to determine whether additional assessments were as accurate as the previous assessments. One study had addressed some of those issues but shared the concern that only one reassessment was checked. This study addressed some of the issues with earlier studies by testing up to three reassessments, using a larger sample that was a larger portion of the total population of offenders, and testing in another country (the U.S.). It shared a concern with some of the previous studies in the use of a non-random sampling method.

This study used data from follow-up LSI-R assessments done with offenders used in the validation portion of this study. The predictive validity of the successive LSI-R assessments was assessed using replications of previous research on the dynamic predictive validity of the LSI-R, graphs to show the changes in LSI-R scores and regression to the mean, measurement of correlation rates, AUC values to compare the predictive validity of the successive LSI-R assessments, and logistic regression analysis to determine whether changes were significant.

The results of the analyses were somewhat mixed. The first set of analyses examined the score changes that occurred from LSI-R #1 to LSI-R #2. There was a marked regression to the mean between assessments, but the differences in predictive validity between the first and second LSI-R were as expected, with a significant increase in predictive ability for LSI-R #2 over LSI-R #1 for both the AUC values and the correlation rates. The logistic regression model showed that the changes in scores from LSI-R #1 to LSI-R #2 were a significant contributor to prediction accuracy.

The second set of analyses examined the score changes that occurred from LSI-R #2 to LSI-R #3. There was some regression to the mean. High scores went down to a greater extent than the low scores went up. LSI-R #3 was a better predictor of violation than LSI-R #1 but the gains in predictive accuracy were much smaller when LSI-R #3 was compared with LSI-R #2, and were not significant for the AUC and correlation tests. The regression model showed that the changes in score level from LSI-R #2 to LSI-R #3 were a significant contributor to prediction accuracy.

The third set of analyses examined the score changes that occurred from LSI-R #3 to LSI-R #4. The regression to the mean effect was still present. It was much larger for the top half of the distribution. LSI-R #4 was a better predictor than LSI-R #1, but was nearly identical in predictive ability to LSI-R #2 and LSI-R #3. The results using the Raynor method, designed to account for regression to the mean, suggested that LSI-R #4 might have been a worse predictor for some offenders than LSI-R #3. The logistic regression model showed that the changes in score level from LSI-R #2 to LSI-R #3 were not significant contributors to prediction accuracy.

Conclusions

The veracity of the second hypothesis, “LSI-R scores from subsequent assessments are more accurate predictors of risk level than LSI-R scores from previous assessments” is unclear. The results found when comparing LSI-R #1 with LSI-R #2 were exactly as expected, but there did not appear to be a significant difference between LSI-R #2, #3, and #4 in predictive accuracy when comparing the AUC values or correlation rates. The method used by Raynor showed improvements between LSI-R #1 and LSI-R #2 and LSI-R #2 and LSI-R #3 but not between LSI-R #3 and LSI-R #4. The logistic regression models showed that score changes were a significant contributor to the model for LSI-R #2 and LSI-R #3, but not for LSI-R #4.

Since a comparison of the changes in predictive validity between LSI-R #1 and #2 had been performed before, and the results found in this analysis replicated the previous results, that outcome was not surprising. The fact that the follow-up assessments did not replicate the previous results is an unexpected result.

The most logical explanation for the results would seem to be that the larger gains in predictive accuracy between LSI-R #1 and LSI-R #2 were the result of the rater getting to know the offender better. This would tend to explain the mixed results since improvements from LSI-R #1 to LSI-R #2, gained from months of working with the offender, would not be expected thereafter, since the probation officers already knew about as much as they possibly could about the offenders from past experience. The smaller or non-existent gains between the subsequent assessments would then indicate the true level of change in offenders between LSI-R assessments.

An alternative explanation might be that the results are due to sampling error. The samples for the follow-up assessments have more high-risk offenders and are clearly non-random selections from the original population. An examination of Figure 1 shows that the offender risk distributions for LSI-R #2, #3, and #4 were all similar. This may be part of the reason the other assessments were similar in predictive ability. There is also a significant central tendency in the selections from LSI-R #2 to #3, and from LSI-R #3 to #4 (see Figures 5 & 7). This central tendency might explain why the results for LSI-R #2, #3, and #4, are similar, since some of the LSI-R scores didn't change much.

The results might have been affected by regression to the mean (RTM). There was some RTM for all three samples, although the largest regression was between LSI-R #1 and #2 (see Figures 4, 6, & 8). James (1973) cites two areas that can cause RTM, day-to-day variations in the person, and variation in measurement. Both could certainly be present in the measurement of LSI-R scores over time. There is some question about the effects of the RTM since the overall predictive accuracy actually increased the most where the RTM was the greatest. This would tend to support the hypothesis that the increase in predictive accuracy was due to a reduction in error level between assessments. Presumably, scores that were extremely high or extremely low would be more likely to be partly due to random errors. The error would presumably be less on the second assessment. The results found bring up some of the issues brought up by Willett (1989; 1994) about the problems with using scores from two wave data.

Study Limitations

Brown (2003) had indicated that research using just a single measure to assess offenders was weak because it gives a limited view of the offender. This study only uses the LSI-R scores to assess risk level and so would be weak by her description.

Cook et al. had indicated that statistical conclusion validity is affected by non-random sampling. Since this study used three consecutive non-random samples, this certainly may have affected the results.

Recommendations for Future Research

Since this is the first study to look at multi-wave results for the LSI-R, replication would need to be done before any conclusions can be drawn. Replication with random samples would be preferred. Further analysis might try to determine whether the length of time between assessments is a factor. Both Bonta and Lowenkamp (personal communications, 2007) indicated that change takes time to occur and follow-up assessments that are made too quickly might not show any change between assessments. This would not explain why changes actually did occur between the first two assessments however. A more in-depth look at the data using the methods proposed by Willett (1989; 1994) could be done to look at how scores change for individuals as well as the group. The RTM effects should be studied to determine whether they are a significant problem when doing reassessments. Further research would seem to be important as the results of this study suggest that multiple assessments give smaller and smaller improvements in overall predictive validity.

REFERENCES

REFERENCES

Andrews, D. A. (1982). *The level of supervisory inventory (LSI): The first follow-up*. Toronto: Ontario Ministry of Correctional Services.

Andrews, D. A., & Bonta, J. (1995). *LSI-R: The level of service inventory-revised user's manual.* Toronto: Multi-Health Systems, Inc.

Andrews, D. A., & Bonta, J. (2003). *The psychology of criminal conduct* (3rd ed.). Cincinnati, OH: Anderson Publishing Co.

Andrews, D. A., & Bonta, J. (2006). *The psychology of criminal conduct* (4th ed.). Cincinnati, OH: Anderson Publishing Co.

Andrews, D. A., Bonta, J., & Wormith, S. J. (2004). *The level of service/case management inventory (LS/CMI)*. Toronto, Canada: Multi-Health Systems.

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

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

Bachman, R., & Schutt, R. K. (2003). *The practice of research in criminology and criminal justice* (2nd ed.). Thousand Oaks, CA: Pine Forge Press.

Baird, S., Heinz, R., & Bemus, B. (1979). *Wisconsin case classification/staff deployment project: A two-year follow-up report*. Madison, WI: Department of Health and Social Services. Bureau of Community Corrections.

Blalock, H. (1972). *Social statistics*, NY: McGraw-Hill, 407.

Blumstein, A., & Cohen, J. (1979). Estimation of individual crime rates from arrest records. *The Journal of Criminal Law and Criminology*, *70*(4), 561-585.

Bonta, J. (1996). Risk–needs assessment and treatment. In A. T. Harland (Ed.), *Choosing correctional options that work: Defining the demand and evaluating the supply* (pp.18–32). Thousand Oaks, CA: Sage.

Bonta, J. (2006). Do we need theory for offender risk assessment? Retrieved on April 3, 2007, from http://www.csc-scc.gc.ca/text/pblct/forum/special/espe_c_e.shtml.

Bonta, J., & Motiuk, L. L. (1987). The diversion of incarcerated offenders to correctional halfway houses. *Journal of Research in Crime and Delinquency*, *24*(4), 320-323.

Brown, S. L. (2003). *The dynamic prediction of criminal recidivism: A three-wave prospective study 1995-2002*. Unpublished doctoral dissertation, Queen's University, Kingston, Ontario, Canada.

Burgess, E. W. (1928). Factors determining success or failure on parole. In A. A. Bruce, E. W. Burgess, J. Landesco, & A.J. Harno (Eds.) *The workings of the indeterminate sentence law and the parole system in Illinois*. Springfield, IL: Illinois Committee on Indeterminate-Sentence Law and Parole.

Cook, T. D., & Campbell, D. T. (1997). *Quasi-experimentation: Design & analysis issues for field settings*. Boston: Houghton Mifflin Company.

Crocker, L., & Algina, J. (1986). *Introduction to classical & modern test theory*. Belmont, CA: Wadsworth Group/Thompson Learning.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297-334.

Davenport, C. B. (1914). *Statistical methods: With special reference to biological variation*. New York: Wiley.

DeVellis, R. F. (2003). *Scale development: Theory and applications* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Flores, A. W., Lowenkamp, C., Holsinger, A. M., & Latessa, E. J. (2006). Predicting outcome with the level of service inventory-revised: The importance of implementation integrity. *Journal of Criminal Justice*, *34*, 523-529.

French, S. A., & Gendreau, P. (2006). Reducing prison misconducts: What works! *Criminal Justice and Behavior, 33*(2), 185-218.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute, 15*, 246-263.

Gendreau, P., Goggin, C., & Little, T. (1996). *Predicting Adult Offender Recidivism: What Works! User Report: 1996-07*, Ottawa: Solicitor General of Canada.

Gendreau, P., Goggin, C., & Smith, P. (2002). Is the PCL-R really the “unparalleled” measure of offender risk?: A lesson in knowledge cumulation. *Criminal Justice and Behavior*, *29*(4), 397-426.

Girard, L., & Wormith, S. 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*(2), 150-181.

Glueck, S., & Glueck, E. (1950). *Unraveling juvenile delinquency*. Cambridge, MA: Harvard University Press.

Gottfredson, S.D., & Moriarty, L.J. (2006). Clinical versus actuarial judgments in criminal justice decisions: Should one replace the other? *Federal Probation, 70*(2). Retrieved on March 31, 2007, from http://www.uscourts.gov/fedprob/September_2006/judgements.html.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*, 29-36.

Hanley, J. A., & McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology, 148*, 839-843.

Hare, R. D. (1980). *The Hare Psychopathy checklist*. North Tonawanda, New York: Multi-Health Systems.

Hare, R. D. (1991). *The Hare Psychopathy checklist-revised*. Toronto, Ontario: Multi-Health Systems.

Hare, R. D. (2003). *The Hare Psychopathy checklist-revised* (2nd ed.). Toronto, Ontario: Multi-Health Systems.

Hart, H. (1923). Predicting parole success. *Journal of Criminal Law and Criminology*, *14*(3), 405-414.

Hoffman, P. B. (1983). Screening for risk: A revised salient factor score (SFS 81). *Journal of Criminal Justice*, *11*(6), 539-547.

Hollin, C. R. (2002). Risk-needs assessment and allocation to offender programmes. In J. McGuire (Ed.). *Offender rehabilitation and treatment: Effective programmes and policies to reduce re-offending*. Chichester: John Wiley and Sons.

Holtfreter, K., Reisig, M. D., & Morash, M. (2004). Poverty, state capital, and recidivism among women offenders. *Criminnology & Public Policy, 3*(2), 185-208.

James, K. E. (1973). Regression toward the mean in uncontrolled clinical studies. *Biometrics, 29*, 121-130.

Jenson, T. D. (1998). *Examining the predictive validity of the level of service inventory-revised (LSI-R).* Unpublished Master's Thesis. Mankato, MN.

Kipling, R. (1892). The elephant's child. In J. Beechcoft (Ed.) *Kipling: A selection of his stories and poems*. Garden City, NY: Doubleday (1956).

Latessa, E. J., & Allen, H. E. (2003). *Corrections in the Community* (3rd ed.). Cincinnati, OH: Anderson Publishing Company.

Lowenkamp, C. T., Holsinger, A. M., Brusman-Lovins, L., & Latessa, E. J. (2004). Assessing the inter-rater agreement of the level of service inventory-revised. *Federal Probation, 68*(3), 34-38.

Lowenkamp, C. T., & Latessa, E. J. (2001). *Validating the level of service inventory-revised in Ohio's community based correctional facilities*. Unpublished Manuscript. Cincinnati, Ohio: University of Cincinnati Division of Criminal Justice.

Megaree, E. I., & Bohn, M. J., Jr. (1979). *Classifying criminal offenders: A new system based on the MMPI.* Thousand Oaks, CA: Sage Publications.

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns or cutting scores. *Psychological Bulletin, 52*(3), 194-216.

Mitchell, O., & MacKenzie, D. L. (2006). The stability and resiliency of self-control in a sample of incarcerated offenders. *Crime & Delinquency, 52*(3), 432-449.

Motiuk, L. L. (1991). *Antecedents and consequences of prison adjustment: A systematic assessment and reassessment approach*. Unpublished doctoral dissertation, Carleton University, Ottawa, Ontario.

Motiuk, L. L., Bonta, J., & Andrews, D. A. (1990). *Dynamic predictive criterion validity in offender assessment*. A paper presented at the Canadian Psychological Association Annual Convention, Ottawa, June, 1990.

O'Keefe, M. L., Klebe, K., & Hromas, S. (1998). *Validation of the level of supervision inventory (LSI) for community based offenders in Colorado: Phase II.* Colorado Department of Corrections.

Peterson, W. W., Birdsall, T. G., & Fox, W. C. (1954). The theory of signal detectibility. *IEEE Transactions on Information Theory, 4*(4), 171-212.

Raynor, Peter (2007). Risk and need assessment in British probation: The contribution of LSI-R. *Psychology, Crime, & Law, 13*(2), 125-138.

Reisig, M. D., Holtfreter, K., & Morash, M. (2006). Assessing recidivism risk across female pathways to crime. *Justice Quarterly, 23*(3), 384-405.

Rice, M. E. & Harris, G. T. (1995). Violent recidivism: Assessing predictive validity. *Journal of Consulting and Clinical Psychology, 63*(5), 737-748.

Sampson, R. J., & Laub, J. H. (1993). *Crime in the making. Pathways and turning points through life*. Cambridge, MA: Harvard University Press.

Sampson, R. J., & Laub, J. H. (2005). A life-course view of the development of crime. *Annals of the American Academy of Political and Social Science, 602*, 12-45.

Schlager, M. D., & Simourd, D. J. (2007). Validity of the level of service inventory-revised (LSI-R) Among African American and Hispanic male offenders. *Criminal Justice and Behavior, 34*(4), 545-554.

Stigler, S. M. (1997). Regression towards the mean, historically considered. *Statistical Methods in Medical Research, 6*, 103-114.

Swets, J. A. (1973). The relative operating characteristic in psychology. *Science, New Series, 182*(4116), 990-1000.

Warner, S. B. (1923). Factors determining parole from the Massachusetts Reformatory. *Journal of Criminal Law and Criminology, 14*, 172-207.

Watkins, M. W. (2000). *An EXCEL program for calculating and graphing the Receiver Operating Characteristic (ROC)* [Computer software]. State College, PA: Ed & Psych Associates.

Whaley, R. B. (2001). The paradoxical relationship between gender inequality and rape: Toward a refined theory. *Gender & Society, 15*(4), 531-555.

Willett, J. B. (1989). Questions and answers in the measurement of change. In Ernest Z. Rothkopf (Ed.), *Review of research in education*, Volume 15. Washington, D.C.: American Education Research Association, 345-422.

Willett, J. B. (1994). Measurement of Change. In T. Husen & T. N. Postlethwaite (Eds.), *International encyclopedia of education* (2nd ed.). Oxford, UK: Elsevier Science Press, 671-678.

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

Yule, G. U. (1919). *Theory of statistics*. London: Griffin.