Brief Report: An Implementation Evaluation of the LSI-R as a Recidivism Risk Assessment Tool in Utah

Kort Prince, Ph.D.
Robert P. Butters, Ph.D.

January 2014

THE UNIVERSITY OF UTAH

Utah Criminal Justice Center

COLLEGE OF SOCIAL WORK
COLLEGE OF SOCIAL & BEHAVIORAL SCIENCES
UTAH COMMISSION ON CRIMINAL & JUVENILE JUSTICE
S.J. QUINNEY COLLEGE OF LAW

Brief Report: An Implementation Evaluation of the LSI-R as a Recidivism Risk Assessment Tool in Utah

Project Background

This document is a summary report of a larger evaluation of the implementation of the LSI-R in Utah. Many of the details of the full document have been omitted in order to provide a succinct version of the evaluation that will be easily disseminated. Though omitted from this brief report, the full evaluation provides a literature review detailing the history of assessment instruments, the evolution of the LSI, and competing tools. It also contains considerably more detail on findings and data caveats. The reader is encouraged to view the full report at: http://ucjc.utah.edu/adult-offenders/evaluation-of-the-lsi-r-as-a-risk-assessment-tool-in-utah.

The research project began with an original goal of examining, validating and comparing the predictive validity of the LSI-R (Level of Service Inventory-Revised; Andrews & Bonta, 1995) and the LS/CMI (Level of Service-Case Management Inventory; a shorter assessment that can be calculated using the items from the LSI-R; Andrews, Bonta, & Wormith, 2004) as recidivism assessment tools in the Utah population. However, preliminary analyses of data provided to the Utah Criminal Justice Center (UCJC) by the Utah Department of Corrections (UDC) revealed problems at the data collection level that precluded an accurate test of the tools' respective predictive validities.

Rather than using inaccurate data in attempting to validate the instruments for use as recidivism risk prediction tools in Utah, the present research necessarily altered focus to examine the extent of the data collection problems resulting from software-level issues. Evaluation of the LS/CMI was, therefore, jettisoned, and the research focused instead on identifying and describing the data problems. The present research also discusses evidence suggesting that the LSI-R tool, and its inherent difficulties in administration, may have contributed to item-level, total-score and risk calculation discrepancies noted below.

Data Analysis

Originally, the UDC provided UCJC with data including demographic variables, offense histories and LSI-R item scores for all individuals beginning probation or released to parole between 1/1/2008 and 12/31/2010 (regardless of when the original crime for which they were convicted was committed). Eventually, and in order to examine data inconsistencies described below, the database was extended to all years in which the LSI had been implemented in Utah, and included data from 2000-2013. The full database included 97,641 assessments. The average number of assessments per person was 3.87 but ranged from one to 24 with a median of four assessments. The majority of the assessments were conducted within the parole population (59.9% relative to 40.1% probation). Parole or probation status was not static, however, and the same individual could be a probationer at one time and a parolee at another time.

UCJC calculated domain and total scores for the LSI-R using a UDC-provided file containing the item-level responses of probationers and parolees meeting the aforementioned eligibility criteria. Analyses (using Area Under the Curve (AUC) procedures) revealed relatively poor recidivism prediction for the overall and domain scores of the LSI-R. As a result of the unexpected findings, three hypotheses were postulated to explain why the LSI-R was not as predictive in Utah as other jurisdictions. These included:

1) The tool is simply not as predictive in the Utah offender population;

2) A data quality problem exists (e.g., data are not entered or recorded as intended by the LSI-R administrators); or
3) Administrative problems exist (e.g., a lack of sufficient training or the LSI-R is simply too difficult to complete with fidelity).

In order to investigate whether hypothesis one was most accurate, the latter two would first need to be demonstrated to be false. If data quality or administrative issues were found to exist, hypothesis one could not be examined to the extent that data do not accurately represent the intent of the LSI-R tool or, alternatively, an administrator/user. The issues of data quality and administrative difficulties were examined separately, but it is important to note that one does not preclude the other. They are not mutually exclusive, and both could be potential causes of poor predictive validity that are unrelated to the true predictive validity of the LSI-R if it were administered and recorded as intended.

Addressing Data Quality as a Potential Cause of Poor Predictive Validity

In order to examine whether data quality factors contributed to the low predictive values for recidivism, three primary analyses were conducted.

1) The first analysis examined the number of invalid assessments due to too many missing items;
2) A second analysis compared total scores calculated by the UDC to total scores calculated by UCJC.
3) A third analysis examined the data for item-level response combinations (i.e., if-then items) that are not allowed per instructions of the LSI-R manual (see below).

Invalid Assessments (Missing Items)

Of the 97,641 total assessments over the 13-year period, 13.7% (13,329) should have been considered invalid because too many items were missing to calculate a meaningful LSI-R score. According to the owners of the LSI tools, Multi-Health Systems (MHS), and the LSI-R manual (Andrews & Bonta, 1995), an assessment becomes invalid if it has more than five missing items. The mean number of missing items was 1.87 per assessment, but 6.1% had over 10 missing items. The policy recommended by the LSI-R developers, when more than five items are missing, is to locate the relevant information before finalizing an assessment (Andrews & Bonta, 1995; Andrews, Bonta, & Wormith, 2004). Because LSI-R total scores (as well as corresponding risk levels) were calculated with many items missing, total scores (and consequentially risk categories) are inevitably lower on average than an indeterminable true score. Utah may want to consider adopting the MHS standards and requiring at least a minimum number of items be entered before submission of the assessment can occur (this is not presently the standard protocol in Utah).

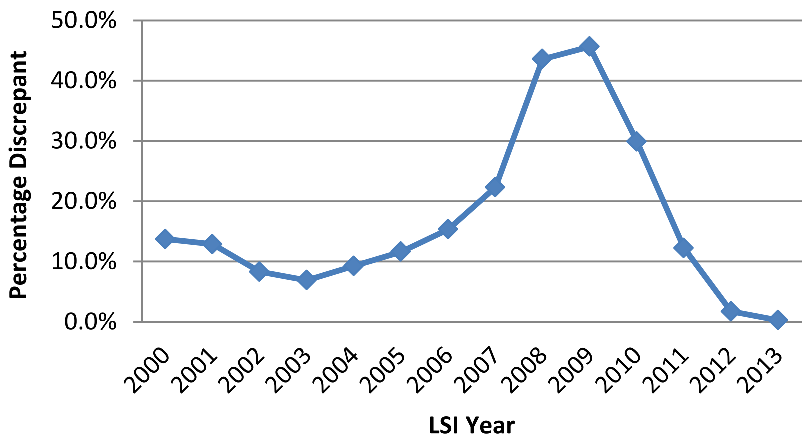
Discrepant Total Scores

A preliminary analysis comparing UDC-provided total scores to UCJC-calculated total scores revealed the UDC scores did not perfectly match UCJC-computed total scores. Furthermore, when the individual item scores from UDC data were summed (the computational method for the LSI-R), those values were different from the UDC total score, indicating that total scores were not a reflection of the item scores within the UDC system. Meetings with UDC staff revealed software-level computational errors were resulting in incorrectly summed total scores (see the full report for more detail).

The research next sought to identify whether the discrepancies were specific to a period of time during which software was updated or replaced. Figure 1 demonstrates that the problem in total score calculation was pervasive since adoption of the LSI-R as an assessment tool; however, the rate at which the

Figure 1

Percentage of Cases With UDC Total Scores that Differ from the UCJC Computed Sum of Item Scores By LSI-R Assessment Year



discrepancies occur begins to drop from 2011 to 2013. The degree of discrepancy fluctuates from year to year, reaching a zenith in 2009 with 45.6% of all assessments conducted in that year not matching the sum of item-level responses. Across the period from 2000-2013, 23.7% of all LSI-R assessments reflected inaccurate total scores. Though 2013 actually shows a small discrepancy (0.3%), the discrepant assessments are limited to the first quarter of 2013. After March of 2013, there were no UDC total scores that differed from the sum of item-level responses, indicating the software-level problem has been fixed. New assessments are correctly totaled, but the software level changes do not negate the inaccuracy of historical assessments (rather, only those conducted after March of 2013).

Because of the problems in total score calculations, mis-categorization of individuals into inaccurate risk levels was also a concern, particularly because assigned risk level dictates the nature and intensity of services and supervision one receives. Table 2 shows the number of individuals at each Utah risk classification level according to the original UDC calculations, and then provides the percentage of individuals who should have been classified at a lower risk level or a higher risk level, and the percentage who were classified correctly despite the total score discrepancy[1]. The table also provides (in parentheses) the LSI-R scores that fall within a classification level as defined by the state of Utah.

Table 2: Percentage of Misclassifications by UDC Original Classification

	True Classification (Sum of Item-Level Responses)			
UDC Classification	**N**	**Lower Level**	**Same Level**	**Higher Level**
Low Risk (0-13)	18,102	NA	99.8%	0.2%
Moderate Risk (14-23)	34,536	8.6%	91.2%	0.2%
High Risk (24-40)	42,917	11.3%	88.6%	0.1%
Intensive (41-53)	2,086	22.3%	77.7%	NA
Overall	97,641	8.5%	91.4%	0.1%

As seen in Table 2, misclassification most often placed a person into a higher risk category than was appropriate based on the true LSI-R total score. The greatest discrepancy in terms of percentage misclassified occurred within the intensive risk category, wherein 22.3% of individuals classified as intensive should have been classified in a lower risk category. Although this was the category with the greatest percentage of misclassified cases, it impacted the smallest number of people (i.e., 22.3% of 2,086). Overall, 8.5% of all LSI-R assessments between

[1] Note that denoting classifications as "correct" is not entirely accurate. As discussed later in the report, administrative errors that led to prohibited item-level response combinations also affect the accuracy of the total score. Here, in the absence of validation checks correcting for impossible item-level response combinations, the sum of the item-level responses is considered the best approximation of the "correct" or "true" LSI-R score.

2000 and 2013 resulted in a classification of an offender into a higher risk than was appropriate (i.e., true classification should have been a "Lower Level"); only 0.1% of all cases were classified into a lower risk level than was appropriate (i.e., true classification should have been a "Higher Level").

Prohibited Combinations of Item-Level Responses

The LSI-R manual and scoring guide both dictate the scoring of items with if-then logic; that is, if one statement is true (or false), then another statement must also be true (or false). Although there are a number of items with dependent if-then logic, for the sake of brevity, this brief report provides only one example (for the if-then dependencies related to employment items (the items for which violations were most frequent)). A detailed listing of the items with if-then logic (and the frequency with which they were violated) can be found in the full report.

The scoring manual for the LSI-R dictates the following scoring logic for if-then employment-related items: If a person is noted as currently unemployed (item 11), then work specific participation/performance (item 18), peer interactions (item 19) and authority interactions (item 20) must also be marked as problems. The prohibited response combination of marking item 11 as problematic (unemployed) without marking items 18-20 as problematic as well occurred 41.6% of the time. To some extent, this scoring rule may seem counterintuitive to the LSI-R administrator, but the intent of marking all of these items as a problem, despite the lack of actual employment, is to weight unemployment more heavily, such that, if one is unemployed, it is a relatively greater factor in determining the overall score[2]. Though this is the psychometric logic behind scoring items 18-20 as problems even when a person does not have a job, the reasoning is not explained in the manual, and is perhaps often misunderstood.

Other violations of if-then logic, though present, were considerably rarer, and ranged from as low as 0.3% of all assessments violating an if-then rule to a (non-employment related item) maximum of 4.4%. Mistakes in if-then logic are both a data quality issue and an administrative issue; however, because they can be stopped at the data-quality level through software-level data validation checks, they have been discussed in the data quality section. Also, it is important to note that, as of this report, data quality checks are present in the UDC system that will not allow submission of an LSI-R assessment if any of the if-then logic is violated. The user will receive an error message upon attempted submission and will be prompted to fix items violating the if-then logic. Total scores are also now calculated correctly and the sum of item-level responses equals the calculated total score in all cases. However, as of this report, a user could still submit an assessment with more than five missing items.

Addressing Administrative Issues as a Potential Cause of Poor Predictive Validity

The current research also examined administrative-level factors as potential causes of lower-than-expected predictive validity for the LSI-R in Utah. These included: (1) inadequate training, (2) inherent difficulty of the LSI-R assessment, and (3) ambiguities on the LSI-R response form.

Inadequate Training

Issues addressed above, under data quality concerns, demonstrate that concerns also exist regarding how the LSI-R is administered and the adequacy of training received. For example, if administrators were fully aware of the if-then logic of many of the items, errors related to prohibited item-level response

[2] A person is not considered unemployed if he or she is working part-time, is a seasonal worker, is in a work skills training program, is a full-time student, is a homemaker or a pensioner, is retired, is employed in the institution in which he or she is incarcerated, or is serving less than two years and will verifiably return to the same job after release.

combinations could be reduced substantially. This outcome suggests a lack of familiarity with the LSI-R manual and, perhaps, a lack of training.

Another issue suggesting a lack of sufficient training and familiarity with the LSI-R’s directions involved the frequency with which certain items were left blank or skipped. While all items were skipped to some degree (range 0.5% missing/skipped to 8.9% missing/skipped), certain items were skipped at a relatively alarming rate (i.e., nearly 1 in 11 assessments skipped the items; see Table 3); moreover, these frequently-skipped items clustered into specific domains of the LSI-R. For example, recall that item 11 assesses current unemployment, and that items 18, 19 and 20 (assessing work-related participation/performance, work peer interactions, and work authority interactions, respectively) must be marked as problems if the individual was unemployed. These four related items were four of the five most commonly skipped items in the LSI-R data. The frequency with which these items were skipped may suggest a lack of familiarity with the LSI-R’s if-then logic. The aforementioned finding that these items often revealed prohibited response combinations augments the possibility that administrators simply do not know how to respond or do not understand the logic of the directions with respect to these items. Particularly with offenders who are, or have recently been, incarcerated, administrators may find these items difficult to complete. However, the LSI-R manual does offer guidelines regarding how to score these items for incarcerated individuals.

Table 3: Ten Most Commonly Skipped Items on the LSI-R

Item	**Skip Frequency (%)**
40. Drug problem, currently?	8.9
18. Work participation/performance	8.6
20. Work authority interactions	8.4
11. Currently unemployed?	8.2
19. Work peer interactions	8.0
41. Law violations?	7.4
31. Could make better use of free time	6.6
21. Financial problems?	6.3
27. Unsatisfactory accommodations	6.1
23. Dissatisfaction with marital or equivalent situation?	5.7

Unfortunately, the present study could not interview LSI-R administrators in order to determine whether they had received proper training or supervision because interviews were not included in the IRB-approved study design (as they were not needed given the original intent of the project); however, a report issued by the Office of the Legislative Auditor General for the State of Utah (2013) addresses the issue of training. Interviews conducted with Adult Probation and Parole (AP&P) staff as part of the audit indicated that “some staff members stated that agents have not been trained on the proper administration of the LSI-R, instead learning on the job from other employees and self-study” (see full report, p. 21). The audit references an independent report (also reviewed as part of the present research) by Lowenkamp, Latessa, and Holsinger (2004) providing evidence that the predictive validity of the LSI-R declines substantially when staff are inadequately trained in its administration. Recall that the LSI-R’s predictive validity in Utah was initially examined as part of the current research project, but the outcomes from those analyses were jettisoned because both data quality and administrative-level errors were found.

Inherent Difficulty of the LSI-R Assessment

The LSI-R is so widely utilized as a risk and needs assessment tool that it is largely unquestioned whether it is actually difficult to administer in practice. However, Austin (2006) raises concerns regarding difficulties administering the LSI-R and warns of the subjective nature of many of the LSI-R items. Regardless of research on the instrument, however, it is easy to see (upon reviewing the actual instrument) how different administrators could have difficulty reliably providing the same responses to an individual’s LSI-R assessment. For example, one can see how it would be difficult for an administrator, interviewing an individual incarcerated for many years, to decide whether the individual has only "some criminal friends." Similarly, the response (yes or no) to whether the individual has ever had an alcohol problem or a drug problem states only: "the assessment of an 'alcohol' (drug) problem depends upon the interviewer's assessment and not the offender's evaluation" (Andrews, Bonta, & Wormith, 2004, p. 18). The criteria for making such an assessment are not further specified.

Some evidence of administrative difficulties/inconsistencies was found in the UDC data. Prohibited item-level response combinations and frequently skipping items may be partially driven by a lack of understanding of the LSI-R manual, but, as the examples above illustrate, even with an adequate knowledge of the manual's scoring guide, some items are simply ambiguous and subject to a degree of speculation and subjective interpretation.

Ambiguities on the LSI-R Response Form

LSI-R administrators complete the assessment on a software platform developed by UDC. A member of UCJC staff reviewed the system with a member of the supervisory staff from UDC and was able to address specific anomalies/concerns found in the LSI-R data during analyses. Details of the review are outlined in the full report, which documents aspects of the UDC web-based system that may contribute to ambiguities in how to respond. For example, unlike the paper version, where questions in a domain are clearly marked by a large header, the web-based version has a small header and, once a user scrolls down on the page, the heading is no longer visible. Consequently, it is not always clear that subsequent questions relate to the previous header/domain.

Conclusion

This research project began with an original goal of examining and validating the LSI-R and LS/CMI as recidivism assessment tools in the Utah population. However, preliminary analyses of LSI-R data provided to UCJC revealed problems that precluded testing the validity of the LSI-R in Utah. Problems included both data quality issues (i.e., invalid tests due to too many missing items, total scores that did not match the sum of item-level responses and prohibited combinations of item-level responses) and administrative issues (inadequate lack of training of administrators, inherent difficulties in using the LSI-R and ambiguities on the LSI-R response form).

The new UDC system (adopted in March of 2013) has eliminated many of the problems that precluded an evaluation of the tools' predictive validities; the newest LSI-R software iteration now correctly calculates total scores, and will not allow prohibited item-level response combinations. It will, however, still allow more than five missing items, a decision that should remain a topic of future concern because missing items invalidate the assessment and artificially lower total scores. Moving forward, other improvements can be made that will facilitate a fair and accurate evaluation of the LSI-R (and its overall usefulness as a risk and needs classification tool). These include: 1) ensuring all users are adequately trained, and 2) improving the ease of interpreting the software's user interface. Difficulties inherent in the LSI-R will remain an important issue for discussion, but much of this difficulty can be attenuated by ensuring minimum standards of training on the instrument. Increased training may have the additional benefit of helping users better understand the instrument, creating greater faith in its validity and overall usefulness.

References

Andrews, D. A., & Bonta, J. (1995). *The Level of Service Inventory–Revised.* Toronto, Ontario, Canada: Multi-Health Systems.

Andrews, D. A., Bonta, J. L., & Wormith, J. S. (2004). *Level of Service/Case Management Inventory (LS/CMI): An offender assessment system,* u*ser's manual*. Toronto, Ontario, Canada: Multi-Health Systems.

Austin, J. (2006). How much risk can we take? The misuse of risk assessment in corrections. *Federal Probation, 70*(2), 58-63.

Office of the Legislative Auditor General State of Utah. (2004). *Performance audit of the division of Adult Probation and Parole: Report to the Utah Legislature Number 2013-08*. Salt Lake City, UT: Author.

Lowenkamp, C. T., Latessa, E. J., & Holsinger, A. M. (2004). Empirical evidence on the importance of training and experience in using the Level of Service Inventory-Revised. *Topics in Community Corrections*, 49-53.