Federal PROBATION

a journal of correctional philosophy and practice

Assessing the Inter-rater Agreement of the Level of Service Inventory Revised

Christopher T. Lowenkamp, Ph.D., University of Cincinnati
Alexander M. Holsinger, Ph.D., University of Missouri Kansas City
Lori Brusman-Lovins, M.S.W., University of Cincinnati
Edward J. Latessa, Ph.D., University of Cincinnati

Introduction

Risk assessment has a long history in corrections. In reviewing the types of assessment practices available, Bonta (1996) identifies three generations of risk assessments. Each of these assessment processes possesses advantages and disadvantages. For example, the first generation of risk assessments, also known as quasi-clinical or subjective assessments, allows for deviation from the assessment protocol when necessary, but has proven to be lacking in predictive accuracy (Bonta, Law, & Hanson, 1998; Hanson & Busiere, 1998; and Mossman, 1994). Second generation assessments are objective and empirically based, but often focus on criminal history and a host of atheoretical (and static) factors. While the second generation of risk assessments has been fairly accurate in regard to prediction and easy to score, very little can be garnered from this second generation that leads to the development of a meaningful intervention plan (Andrews & Bonta, 2003). The third generation of risk assessments is also objective and empirically based. What makes them more useful for developing case planning is their dynamic measurement of risk factors and the quality and breadth of information collected. But this advantage is also a disadvantage. Measuring dynamic risk factors and scoring a detailed and comprehensive risk assessment requires specialized knowledge of the assessment process and the items contained therein.

The advent of the third generation of risk assessments has provided much hope for correctional interventions. Such assessments can be used not only to identify high-risk offenders, but to determine what factors exist in an individual's life that cause him or her to be high risk (Lowenkamp & Latessa, 2002). Such a determination provides meaningful targets for interventions. If these targets are properly addressed, reduction in risk and subsequently the likelihood of recidivism follow. In the aggregate, these instruments can assist correctional agencies in increasing public safety. However, as noted earlier, completing these assessments requires training and considerably more care than completing other assessment methods. The LSI-R, the focus of the current research, requires knowledge of psychological testing in general and specialized knowledge of how to score the risk assessment itself (Andrews & Bonta, 2001). Included in the LSI-R are 10 criminogenic domains (criminal history, education/employment,

financial, family relationships, accommodations, leisure and recreation, companions, substance use, emotional health, and attitudes/orientations). Additional reviews of other risk assessments and staff ability to complete and integrate the information in daily decision-making activities also leads to the recommendation of staff training and development (Andrews & Bonta 2003).

One concern, among others, that makes training necessary is the reliability of the completed assessment (for a more complete description of potential threats to the utility of risk assessments, see Bonta, Bogue, Crowley and Motiuk, 2001). With any interview-based, dynamic offender assessment process, reliability in scoring is essential. Often the issue of individual subjectivity is raised due to the ways in which information is gathered, and the scoring criteria that accompany third generation risk/need assessment instruments. One of the major advantages of instruments such as the LSI-R is the potential for standardization in classification and assessment. In other words, when conducted properly, use of the LSI-R may help reduce bias in decision making, create a logical classification strategy, and offer information that can be used to create detailed, dynamic case planning. In light of the weight of the decisions that can be informed through using the LSI-R, inter-rater reliability becomes a critical issue.

To date, much of the research involving the LSI-R has focused on the predictive validity of the tool (Andrews, 1982; Andrews, and Robinson, 1984; Bonta and Andrews, 1993). The LSI-R, through the provision of a composite additive score, should offer a valid scale where high scores are associated with a high probability of recidivism. Conversely, low scores on the LSI-R should represent a low probability of recidivism. The validity of the LSI-R has been shown in a variety of correctional settings and with a variety of offender sub-groups (Lowenkamp and Latessa, 2002; Lowenkamp, Holsinger, and Latessa, 2001). More research is needed, however, regarding the reliability of the LSI-R scores across various raters.

Some studies have indirectly examined the inter-rater reliability of the LSI-R. One study utilized a Self Report Inventory that was derived from the LSI-R itself. The Self Report Inventory was designed to gather from offenders themselves measures similar to those on the LSI-R. The Self Report Inventory did demonstrate inter-rater reliability with the LSI-R, which would indicate congruence between the information gathered by correctional professionals and the information provided by offenders themselves (Motiuk, Motiuk and Bonta 1992). While these results are encouraging, they do not demonstrate reliability across the group of professionals conducting the LSI-R, but rather reliability between the offender being assessed and the professional conducting the assessment. In addition, other research has demonstrated the reliability of the LSI-R, compared to the reliability of other risk assessment tools such as the PCL-R (Gendreau, Goggin, and Smith 2002). While clearly this type of investigation is a necessary part of the correctional research landscape, the question regarding LSI-R scores across individual raters is left largely unanswered. In fact, much of the information about the reliability of the tool takes the form of measures of internal consistency (such as Chronbach's Alpha coefficient) for the 10 subscales present in the tool, as well as the tool as a whole (all 54 items together).

One way to increase inter-rater reliability within the LSI-R (as well as overall quality) may be through staff training (Flores, Lowenkamp, Holsinger, and Latessa 2004). The important effect of training has been demonstrated in other venues as well, involving offender assessments other than the LSI-R (Baird and Prestine, 1988). In order to further test the specific inter-rater reliability of the composite LSI-R score, research that uses a common example across a group of LSI-R raters is necessary. The current research utilizes a sample of correctional professionals, all of whom were formally trained in the use and implementation of the LSI-R. As part of the formal training, a common example was conveyed represented by a vignette containing current (dynamic) narrative information about a particular offender. After the training was complete, the participants were asked to utilize the common vignette to score the 54 items on the LSI-R. The ratings that resulted were used to test the inter-rater reliability of the LSI-R for a sample of trained correctional professionals. The results of the analyses are presented below.

There are several different methods for assessing an instrument's reliability. The focus in this research is inter-rater agreement or the extent to which independent raters converge in terms of their scoring of the same offender. To assess inter-rater reliability, 167 training participants

independently completed an assessment of an offender vignette at the conclusion of a three-day training on the principles of offender classification and the use of the LSI-R in particular.

Methodology

Participants

The participants in this study are 167 correctional practitioners from a large Western state. While data on the individual participants was not collected, participants included males and females, individuals of various races, and those working with offenders in the community and in institutions.

Procedures

The participants in this training were part of a three-day training required of all correctional staff. The training covered the intent and scoring criterion for each of the 54 items on the LSI-R. At the end of the training, the participants were given an exam that included a vignette describing an offender. This vignette covered all areas represented on the LSI-R. Participants were instructed to complete the exam and score an LSI-R based on the vignette independently. A facilitator was present during the completion of the exam. The scores from the LSI-R scoring forms were entered into a database for analyses.

Analyses

Since there was only one assessment for each of the 167 raters, traditional tests of inter-rater reliability are not possible. However, investigating the percentage of agreement for each item, descriptives for the total score, and agreement for overall classification are possible. To assess the reliability of the LSIR, the percentage of raters that scored each LSI-R item was calculated. Marks for each item were coded as either indicating a risk factor, indicating that the item was not a risk factor, circled items, and items that were left blank. Average agreement percentages were calculated for each section and for the entire instrument. Descriptive statistics were calculated on the overall score and the final classification based on that score.

Results

Table 1 presents by-item results for the first four sections of the LSI-R (Criminal History, Education/Employment, Financial, and Family/Marital). The percentage of the respondents who scored each item in a particular way is presented. There were four possibilities for each item: 1) marking an item as a risk factor, 2) marking an item as a non-risk factor, 3) circling an item, indicating that not enough information was present to assess it either way, and 4) leaving the item blank. For the Criminal History section, the agreement was very high. For nine of the ten items, agreement ranged from 86 percent to 100 percent. The lowest agreement occurred for "ever punished for institutional misconduct," where 55 percent of the sample scored the item as a non-risk factor. All the items within the Education and Employment section had a percentage agreement ranging between 95 percent and 99 percent. The Financial section, with only two items, had fairly low agreement by comparison, with 66 percent and 57 percent of the sample in agreement. Three of the four items in the Family and Marital section had very high agreement, ranging between 94 percent and 100 percent. The last item in this section had an agreement rate of 51 percent.

Table 2 presents by-item results for the remaining six sections of the LSI-R (Accommodation, Leisure/Recreation, Companions, Alcohol/Drug, Emotional/Personal, and Attitudes and Orientations). The three items in the Accommodations section had very high rates of agreement (96 to 99 percent).

Similarly, the Leisure and Recreation section also showed high rates of agreement for the two items in the section (90 and 98 percent). Likewise, the five items in the Companions section had high rates of agreement (89 to 100 percent). The first seven (of nine) items in the Alcohol and Drug section had very high rates of agreement amongst the raters (92 to 100 percent). However, the last two items had agreement rates of lesser magnitude (56 and 72 percent). Four of the five items in the Emotional and Personal section had high rates of agreement, ranging from 89 to 99 percent. One item had a moderate rate of agreement, at 65 percent. Finally, the Attitudes and Orientations section showed high rates of agreement as well, with the four items ranging in agreement from 82 to 98 percent. Overall, the agreement rate for the 54 items taken as a whole was very high for the sample.

Table 3 presents the average agreement rates for each subsection. Nine of the 10 subsections had average agreement rates of 85 percent or above (the Accommodations section with three items had the highest average agreement rate at nearly 98 percent). The Financial section, however, had the lowest average agreement rate, at 61.5 percent.

Table 4 presents the portions of the sample that placed the offender in each category of risk (using Multi-Health System's prescribed cut-off scores). A large majority of the subjects in the sample—86 percent—assessed the offender in the vignette as having a composite score that placed them into the Medium/High category of risk. These results are particularly important when considering the importance of classifying offenders objectively, and allowing agencies to incorporate the Risk principle of correctional classification and intervention. Regardless of slight differences that may have occurred across a handful of the items across raters, overall, the average rates of agreement were acceptable to very high for each subsection, and a very high proportion of the sample were assessing the offender as being at the same level of risk.

Discussion

The goals of the current research were fairly modest. After being trained on the use of the LSI-R, practitioners were tested as to whether or not they agreed on the scoring of the 54 items present on the assessment. However modest these goals were, the process of determining whether or not practitioners can reliably score the LSI-R assessment is an important issue. The LSI-R and the information it gleans can be used to inform several decision points throughout the processing of offenders. In addition, both the Risk and Need principles can be met via the use of the LSI-R. In light of these aspects of offender assessment and classification, and the fact that the LSI-R is an example of a proprietary tool that requires agencies to commit resources, inter-rater reliability becomes even more important. Based on the results presented above, properly trained practitioners do exhibit high levels of agreement across virtually all the items in the 54-point scale. Even considering that a small number of items had moderate rates of agreement, the overall average agreement rates for all 10 subsections were acceptable to very high. In most cases, where agreement rates were lower, that should be interpreted in light of the fact that the subjects had only attempted to conduct the assessment two prior times. As such, with continued practice and quality assurance checks, rater agreement should only increase over time. This assumption is supported by Flores et al. (2004), who found that the amount of time an agency uses the instrument and the implementation of formalized training on its use produce significant increases in the predictive validity of the tool.

Some limitations were inherent in the current research. For example, the LSI-R process, when conducted in the field, requires practitioners to gather their own data via one-on-one interviews with offenders and the consideration of multiple sources of collateral information. The subjects in the current study were given a tailor-made vignette that represented the information that should have been gathered had they been involved in a real-life assessment process. A true test of inter-rater reliability using the LSI-R would require pairs (or more) of subjects to gather their own information independently from the same source, after which the assessment would be scored accordingly. Doing this would also allow for the calculation of inferential statistics designed to more explicitly test rater agreement. Nonetheless, the research presented above represents a

descriptive analysis that attempts to contribute to the knowledge base pertaining to inter-rater reliability with practitioners who have been trained in the use of the LSI-R.

References

Item	Percent Risk Factor	Percent Not a Risk Factor	Percent Circling Item	Percent Left Blank
Any Prior Adult Convictions	99	0	0	1
Two or more prior convictions	98	2	0	0
Three or more prior convictions	98	2	0	0
Three or more present offenses	3	95	1	2
Arrested under age 16	96	4	0	0
Ever incarcerated upon conviction	99	0	0	1
Escape history from a correctional facility	1	86	13	1
Ever punished for institutional misconduct	15	55	27	3
Charge laid or probation/parole suspended during prior community supervision	99	1	0	0
Official record assault or violence	0	100	0	0
Unemployed	97	2	0	0
Frequently unemployed	99	1	0	0
Never employed for a full year	95	5	0	0
Ever fired	99	1	0	0
Less than regular grade 10	1	99	0	0
Less than regular grade 12	1	99	0	0
Suspended or expelled at least once	2	98	0	0
Participation/performance	99	1	0	0
Peer interactions	95	4	2	0
Authority interactions	95	2	2	1
Financial problems	66	2	22	10
Reliance upon social assistance	7	23	57	13
Dissatisfaction with marital or equivalent situation	100	0	0	0
Nonrewarding, parental	96	4	0	0
Nonrewarding, other relatives	94	6	0	0
Criminal family/spouse	9	51	37	3

Table 1: Agreement Percentages for LSI-R Subsections Criminal History, Education/Employment, and Financial

Item	Percent Risk Factor	Percent Not a Risk Factor	Percent Circling Item	Percent Left Blank
Unsatisfactory	99	0	0	1
Three or more address changes last year	96	4	0	0
High crime neighborhood	98	1	1	0
Absence of a recent participation in an organized activity	90	8	1	1
Could make better use of time	98	1	1	1
A social isolate	92	8	0	1
Some criminal acquaintances	99	1	0	0
Some criminal friends	100	0	0	0
Few anticriminal acquaintances	89	9	0	2
Few anticriminal friends	95	5	0	1
Alcohol problem, ever	99	1	0	0
Drug problem, ever	100	0	0	0
Alcohol problem, currently	96	4	0	1
Drug problem, currently	95	4	0	1
Law violations	99	1	0	0
Marital/family	99	3	1	1
School/work	95	5	2	1
Medical	13	56	29	2
Other indicators	72	15	6	7
Moderate interference	97	3	0	0
Severe interference	35	65	0	0
Mental health treatment, past	1	99	0	0
Mental health treatment, present	2	98	1	0
Psychological assessment indicated	89	10	1	1
Supportive of crime	17	82	1	0
Unfavorable toward convention	7	93	0	0
Poor toward sentence	2	98	0	1

Poor toward supervision	2	98	0	0

Table 3: Average Percentage Agreement by LSI-R Subsection		
Area	**Average Agreement**	**Number of Items**
Criminal History	92.50	10
Education/Employment	97.50	10
Financial	61.50	2
Family/Marital	85.25	4
Accommodations	97.66	3
Leisure/Recreation	94.00	2
Companions	95.00	5
Alcohol/Drug Problems	89.77	9
Emotional/Personal	89.60	5
Attitudes/Orientation	92.75	4

Table 4: Risk Level Classification

Classification Level	Percent	Number
Low/Moderate	1	2
Moderate	11	19
Medium/High	86	143
High	2	3

Peluso, E. and N. Putnam. (1996). Case study: Sexual abuse of boys by females. *Journal of the American Academy of Adolescent Psychiatry,* 35(1), 51–54.

Rosencrans, B. (1997). *The Last Secret: Daughters Sexually Abused by Mothers.* Orwell, VT: Safer Society Press.

Sack, William H. and Robert Mason. 1980. Child abuse and conviction of sexual crimes: A preliminary finding. *Law and Human Behavior,* 4, (3): 211–215.

Szymkowiak, K., & Fraser, T. (2002). *Registered Sex Offenders in Hawaii.* Honolulu: Department of the Hawaii Attorney General.

Tewksbury, R. (2002). Validity and utility of the Kentucky sex offender registry. *Federal Probation,* 66(1), 21–26.

Tewksbury, R. (in press). Collateral consequences of sex offender registration. J*ournal of Contemporary Criminal Justice.*

U.S. Department of Justice, 2002. *Crime in the United States, 2001.* Washington, D.C.: Federal Bureau of Investigation

Vandiver, D. M., & Walker, J. T. (2002). Female sex offenders: An overview and analysis of 40 cases. *Criminal Justice Review,* 27(2), 284–300.

Assessing the Inter-rater Agreement of the Level of Service Inventory Revised

Andrews, D.A. (1982). *The Level of Supervision Inventory (LSI): The First Follow-Up.* Toronto, Ontario: Ontario Ministry of Correctional Services.

Andrews, D.A. & J. Bonta (2003). *Psychology of Criminal Conduct.* Cincinnati, OH: Anderson Publishing Co.

Andrews, D.A. & J. Bonta (2001). *LSI-R User's Manual.* Toronto, Ontario: Multi-Health Systems, Inc.

Andrews, D.A. & D. Robinson (1984). *The Level of Supervision Inventory: Second Report.* Toronto, Ontario: Ontario Ministry of Correctional Services.

Baird, C. & R. Prestine (1988). *Revalidation of the Illinois Risk Assessment System.* South Carolina: South Carolina Office of the Governor.

Bonta, J. (1996). Risk-Needs Assessment and Treatment. In A. T. Harland (Ed.), *Choosing Correctional Interventions That Work: Defining the Demand and Evaluating the Supply.* Newbury Park, CA: Sage, pp. 18-32.

Bonta, J., B. Bogue, M. Crowley, M., & L. Motiuk (2001). Implementing offender classification systems: Lessons learned. In G. A. Bernfeld, D. P. Farrington & A. Leschied (Eds.), *Offender Rehabilitation in Practice* (pp. 227–245). Chichester, England: Wiley & Sons.

Bonta, J., M. Law, and K. Hanson. (1998). The Prediction of Criminal and Violent Recidivism Among Mentally Disordered Offenders: A Meta-Analysis. *Psychological Bulletin,* 123(2), 123–142.

Bonta, J. & D.A. Andrews (1993). The Level of Supervision Inventory: An Overview. *IARCA Journal* 5(4):6–8.

Flores, A.W., C.T. Lowenkamp, A.H. Holsinger, & E.J. Latessa (2004). *Predicting Outcome With the Level of Service Inventory—Revised: The Importance of Quality Assurance.* Unpublished manuscript.

Gendreau, P., C. Goggin, & P. Smith (2002). Is the PCL-R Really the "Unparalleled" Measure of Offender Risk? A Lesson in Knowledge Accumulation. *Criminal Justice and Behavior,* 29(4):397–426.

Hanson, R.K. & M.T. Bussiere (1998). Predicting Relapse: A Meta-analysis of sexual offender recidivism studies. *Journal of Consulting and Clinical Psychology,* 66: 348–362.

Lowenkamp, C.T. & E.J. Latessa (2002). Norming and Validating the LSI-R: Northwestern State Sample. University of Cincinnati: Unpublished manuscript.

Lowenkamp, C.T., A.H. Holsinger, & E.J. Latessa (2001). Risk/Need Assessment, Offender Classification, and the Role of Childhood Abuse. *Criminal Justice and Behavior,* 28(5):543–563.

Mossman, D. (1994). Assessing Predictors of Violence: Being Accurate About Accuracy. *Journal of Consulting and Clinical Psychology,* 62:783–792.

Motiuk, M.S., L.L. Motiuk & J. Bonta (1992). Comparison Between Self-Report and Interview-Based Inventories in Offender Classification. *Criminal Justice and Behavior,* 19(2):143–159.

Planning for the Future of Juvenile Justice

Allen, Francis A., *The Borderland of Criminal Justice,* Chicago: The University of Chicago Press, 1964.

Bank, L., C. Reddick, G. Weeber, and A. Swinehart, "Motivation and Parenting with an Offender Population," *Community Corrections Report* 11(May–June, 2004) 4.

Blau, P.M. and W.R. Scott, *Formal Organizations,* San Francisco: Chandler, 1962

Boguslaw, R., *The New Utopians: A Study of System Design and Social Change,* Englewood Cliffs, NJ: Prentice-Hall, 1965.

Cohn, Alvin W., "Training in the Criminal Justice Nonsystem," *Federal Probation,* June, 1974.

Cohn, Alvin W., (ed.) *Criminal Justice Planning and Development,* Beverly Hills: Sage Publications, 1977.

Dror, Y. "The Planning Process: A Facet Design," in *Planning, Programming, Budgeting: A Systems Approach to Management,* Chicago: Markham, 1968.

Ewing, D.W., *The Practice of Planning,* New York: Harper and Row, 1968.

Follett, Mary Parker, "The Process of Control," in L. Gulick and L. Cambell (eds.), *Papers on the Science of Administration,* New York: Institute of Public Administration, 1937.

Freed, Daniel J., "The Nonsystem of Criminal Justice," in J.S. Campbell, J.R. Sahid, and D.P. Stang (eds.), *Law and Order Reconsidered: A Staff Report to the National Commission on the Causes and Prevention of Violence,* Washington, D.C.: U.S. Government Printing Office, 1969.

Hambrick, D.C., M.A. Geletkanycz, and J. W. Fredrickson, "Top Executive Commitment to the Status Quo: A Test of Some of Its Determinants," *Strategic Management Journal,* 14 (1993).

Higham, John, *Strangers in the Land,* New Brunswick, NJ: Rutgers University Press, 1963.

Katz, Michael S., "Criminal Justice Planning: Prediction or Projection," in Cohn, A.W. (ed.),