CORRECTIONS: POLICY, PRACTICE AND RESEARCH
https://doi.org/10.1080/23774657.2017.1323253

An Examination of the Inter-Rater Reliability and Rater Accuracy of the Level of Service/Case Management Inventory

Ryan M. Labrecque[a], Christopher M. Campbell[a], Jaycee Elliott[a], Megan King[a], Molly Christmann[a], Kari Page[a], John McVay[b] and Katie Roller[b]

[a]Department of Criminology and Criminal Justice, Portland State University, Portland, Oregon, USA; [b]Multnomah County Department of Community Justice, Portland, Oregon, USA

ABSTRACT

Correctional agencies use risk assessment instruments for a wide range of purposes, including to help classify, manage, and treat offenders. The literature on offender risk assessment largely focuses on assessing for predictive accuracy, and far less research examines reliability in scoring. This study adds to this gap in knowledge by assessing how reliably and accurately a group of trained raters score one particular risk assessment tool, the Level of Service/Case Management Inventory (LS/CMI). Findings reveal an adequate to strong level of inter-rater reliability across the domains of the LS/CMI. These results also suggest there is a wide range of rater accuracy across the items and domains of the LS/CMI. The policy and practical implications of these findings are discussed.

KEYWORDS

Risk assessment; reliability; community corrections; probation

Over the last three decades, correctional agencies have adopted a wide array of risk assessment instruments to help classify, manage, and treat offenders (Latessa & Lovins, 2010). The intent of these instruments is to help justice officials identify which offenders are more likely to engage in a variety of unwanted behaviors, including recidivism, institutional misconduct, and failure to appear in court (Andrews, Bonta, & Wormith, 2006). It is important that these tools are capable of effectively differentiating between offenders based upon their probability for engaging in these different types of problematic outcomes (Austin, 2006). The accurate assessment of risk is critical because this information can be used to identify the characteristics or aspects of the offender and his or her situation that may be targeted with intervention to reduce their risk of failure (Bonta, 2002). Moreover, it is also important that justice officials consistently and accurately evaluate offender risk to reduce error in making decisions related to personal liberty and public safety, including choices to sentence, release, revoke, and how intensely to supervise (Singh, Fazel, Gueorguiva, & Buchanan, 2014).

Examinations of the predictive accuracy of risk instruments are plentiful (see e.g., Taxman, 2017). However, among the many studies in this area, the focus of predictive validity is almost exclusive, giving little attention to concerns of reliability. To some this may not seem problematic. After all, several instruments consist mainly of objective, static measures (e.g., criminal history items), which maintain high reliability as well as predictive validity. On the other hand, an increasing number of risk assessments contain items that are more subjective. Instruments that aim to capture measures of dynamic risk often require a certain level of rater judgment

CONTACT Ryan M. Labrecque rml@pdx.edu Assistant Professor, Portland State University, Department of Criminology and Criminal Justice, 506 SW Mill St., PO Box 751, Portland OR 97207, USA.

(e.g., items related to an offenders' attitude or personality). Thus, it is imperative that risk instruments be assessed not only for predictive accuracy, but also for inter-rater reliability (i.e., the degree of agreement among raters).

This study seeks to advance scholarship on offender risk assessments by addressing this gap in knowledge. More specifically, it assesses how reliably a group of trained raters in a Pacific Northwest urban university setting score the items and domains of one risk assessment tool: The Level of Service/Case Management Inventory (LS/CMI; Andrews, Bonta, & Wormith, 2004). Additionally, it examines how consistent or accurate the trained raters are in relation to the expert trainers' assessment of the same cases. We begin with a review of the prior research informing this work and then discuss the importance of reliability analyses in offender risk assessment. Finally, the study's design, findings, and implications for research and practice are discussed in relation to the context of how risk assessments are implemented in practice.

The principles of effective intervention

The principles of effective intervention (PEI) have become a widely relied upon in offender supervision and rehabilitation (Bonta & Andrews, 2017; Cullen & Gendreau, 2001; Gendreau, 1996). The development of the PEI is the result of an ongoing effort to accumulate knowledge about "what works" to reduce recidivism (Smith, 2013). At the core of PEI are the principles of risk, need, and responsivity (RNR), which suggest that the greatest reductions in recidivism will occur when correctional agencies target the specific criminogenic needs (need principle) of the highest risk offenders (risk principle) with cognitive-behavioral interventions in a manner that are responsive to the offenders individual learning style, motivation, and abilities (responsivity principle; Andrews, Bonta, & Hoge, 1990). A large body of literature finds that those correctional interventions which conform to the PEI produce much greater reductions in recidivism than those that do not adhere to its principles (e.g., Andrews & Dowden, 2006; Andrews, Zinger, et al., 1990; Lowenkamp, Latessa, & Holsinger, 2006). To illustrate, a recent meta-analysis found that programs that adhere to the RNR principles had a 26 percentage point reduction in recidivism, compared to a slight increase in those that did not adhere to any of the principles (Bonta & Andrews, 2017, p. 230; see also Andrews & Dowden, 2006; Andrews, Zinger, et al., 1990; Lowenkamp, Latessa, & Holsinger, 2006).

For interventions to be consistent with the PEI, correctional agencies must first begin with knowledge regarding the offenders' level of risk and criminogenic need (Bonta & Andrews, 2017). Risk assessments subsequently serve as a guide for justice officials to determine how intensively to supervise each offender and which criminogenic need factors to target with interventions (Bonta, 2002; Hollin, 2002). A number of Level of Service (LS) instruments have been developed from the PEI perspective, including the original Level of Service Inventory (LSI; Andrews, 1982), its first revision, the Level of Service Inventory–Revised (Andrews & Bonta, 1995), its juvenile edition, the Youth Level of Service/Case Management Inventory (YLS/CMI; Hoge & Andrews, 2002); and its most recent version, the Level of Service/Case Management Inventory (LS/CMI; Andrews et al., 2004). The LS instruments are among the most widely used and studied offender risk-assessment instruments in North America and abroad (Wormith, 2011).

The LS/CMI is a "fourth generation" tool that integrates case management strategies with the results of the risk/needs assessment (Andrews et al., 2006).[1] The LS/CMI includes 43 items that fall into what are referred to as the "central eight" domains of the PEI (Andrews, Bonta, et al, 1990): criminal history (eight items), education/employment (nine items), family/marital (four items), leisure/recreation (two items), companions (four items), alcohol/drug problems (eight items), procriminal attitude/orientation (four items), and antisocial pattern (four items). Each of these items are scored by a trained rater as either *present* (1 point) or *absent* (0 points) on the basis of a file review, an interview with the offender, and other collateral information that is available (i.e., information learned from interactions beyond the assessment interview). These scores are then summed to construct eight subcategory scores and one overall total risk score. This information is then used to determine the offender's likelihood of recidivism, and to identify which domain(s) are most important to target for intervention to reduce such risk. Although research largely supports the predictive validity of the domains and total score of the LS/CMI on a number of criminal outcomes (e.g., Andrews et al., 2012; Girard & Wormith, 2004; Olver, Stockdale, & Wormith, 2014; Wormith, Hogg, & Guzzo, 2012, 2015), there remains a notable lack of research on the tool's reliability.

The need for more reliability research in offender risk assessment

Some scholars have raised concerns that offender risk assessments may not always be scored by raters in a consistent manner (e.g., Austin, 2006; Baird, 2009). This is problematic because low reliability in scoring will likely decrease an instrument's predictive accuracy (Duwe & Rocque, 2017). This could have far-reaching consequences as assessment results are used in a wide range of decisions that affect offenders' lives (Andrews & Dowden, 2007). Therefore, it is critical to ensure a high level of inter-rater reliability in any risk assessment used to make sentencing, supervision, and treatment decisions (see Cohen, 2017). Curiously, however, reliability remains a largely neglected area of study in the risk assessment literature (Desmarais & Singh, 2013). To illustrate this point, in a recent handbook on offender risk and need assessment published by the American Society of Criminology's Division of Sentencing and Corrections, inter-rater reliability was neither the subject of any chapter, nor was the concept mentioned in the glossary of terms (see Taxman, 2017).

Some evaluations of correctional risk assessments suggest that instruments possess excellent levels of inter-rater reliability, whereas others conclude that tools have less than adequate inter-rater reliability (Cohen, 2017). With respect to the LS instruments more specifically, there are few reliability investigations currently available, and the findings are somewhat mixed. Further, some methodological shortcomings in these works give cause for further study in this area. For example, Austin, Coleman, Peyton, and Johnson (2003) conducted a study of the LSI-R in Pennsylvania and concluded that the tool had poor reliability. More specifically, Austin et al. reported that approximately 60% of the offenders were scored by two raters to have a difference of more than three points. Although it may seem like a negligible number, especially on a scale of 54, it is important to note that for some cases this difference could result in being classified as a different risk level for recidivism (e.g., high risk vs. moderate risk). Indeed, Austin et al. further found that nearly 30% of the offenders were assigned a different level of risk by the two raters. Although these findings raise some questions about the inter-rater reliability of the LSI-R, this study included only two raters and is thus highly sensitive to their

individual abilities. Further, the two assessments were administered 2 months apart from one another, which raises questions about the potential influence of a test–retest effect, and whether or not the differences found were a result of error in assessment or a real change in offender risk (see Andrews, Bonta, & Wormith, 2010).

In another LSI-R reliability study in a large western state, Lowenkamp, Holsinger, Brusman-Lovins, and Latessa (2004) found that the tool had an acceptable level of inter-rater reliability for most domains, with nine of the 10 subsections having an average agreement of greater than 80%. Although this study included 167 correctional practitioners, it was limited to the review of just one case. Likewise, it is not clear if the results were driven by the unique attributes of this one particular offender. Another shortcoming, similar to the Austin et al. (2003) study, is that the authors relied exclusively on the use of percent agreement between raters to assess reliability and did not calculate statistics that correct for the likelihood that raters might agree by chance, such as the intra-class correlation (ICC) coefficient (see Hallgren, 2012).

A YLS/CMI study conducted in Ontario, Canada, by Schmidt, Hoge, and Gomes (2005) reported that the instrument had a good level of inter-rater reliability on the seven subscales examined (offense history was excluded from the investigation). This study, however, provided no information on how the 29 juvenile offenders were selected, how many raters were used, or how the ICC values were computed. Another YLS/CMI study conducted by Baird et al. (2013) concluded that the instrument had a moderate to almost perfect level of reliability across three study locations. This investigation was comprehensive and included a number of reliability measures (e.g., percent agreement and ICC); however, it was also limited to just an examination of the reliability in the tools overall assessment score and did not separate findings by domain.

Most recently, Rocque and Plummer-Beale (2014) conducted a reliability study in a Northeastern state that included the calculation of the ICC for the overall and domain scores of the LSI-R and YLS/CMI. Rocque and Plummer-Beale concluded that both instruments had an adequate to fair level of reliability. However, their findings were separated by domain rather than by item, which leaves some additional question of how reliably individual items within these domains are scored. Further, this study only addressed whether the raters scored the offenders similarly to one another and not whether the scores were correct. This is an important limitation because raters may be reliable in their assessment of offenders, but they may also be consistently inaccurate in their scoring. Low reliability and inaccurate scoring will contribute to a reduction in the predictive validity of the risk assessment tool and could lead to the selection of inappropriate treatment targets.[2]

Current study

This review of the literature highlights the fact that the extent to which offender risk-assessment instruments, generally, and the LS instruments, specifically, are scored reliably is simply unclear. The authors are not aware of any published inter-rater reliability study on the LS/CMI using an ICC, which is especially troublesome because many correctional agencies have adopted this instrument as the most recent and advanced of the LS tools. Thus, the purpose of this study is to assess how reliably and accurately a group of trained raters in a Pacific Northwest urban university setting score the LS/CMI.

Method

Sample and procedure

This study involves the collaboration between faculty and students from the Portland State University (PSU) Department of Criminology and Criminal Justice (CCJ) and training staff from Multnomah County Department of Community Justice (DCJ) in Portland, Oregon. More specifically, three graduate students and one undergraduate senior were approached and selected by the two CCJ faculty for participation in this study. As part of this project, these four students were trained by DCJ staff on the LS/CMI alongside newly hired probation officers. The two training staff are certified as LS/CMI trainers and provide trainings for correctional agencies across Oregon. The initial training included two 8-hour days of instruction on the general background of offender risk assessment and the scoring of the LS/CMI. In addition, about a week later each student individually watched a trainer conduct a LS/CMI interview with an offender in the county's intake unit, and the two parties independently rated the offender. After the interview, the trainer reviewed the student's assessment results and discussed any discrepancies in scoring. Finally, about a month after the initial training, students collectively met with the training staff to watch two videos of officers conducting an LS/CMI assessment with an offender. Students then independently assessed the offenders in the videos with the LS/CMI and were evaluated on their ability to score the items correctly. It should be noted that these were the same videos used by the state to recertify its officers on the LS/CMI. All four students passed all of these requirements, including independently scoring the training videos accurately, which certify them by the state standards as qualified to administer the LS/CMI.

Next, the students were given audio-recordings of adult probationer interviews to score independently. The DCJ training staff provided nine audiotapes that were recorded by county probation officers as part of the application process to become a LS/CMI trainer. The two trainers reviewed all of the audios and collectively agreed upon the appropriate scores for the 43 items of the LS/CMI given the information available. According to this collective assessment, the mean LS/CMI score for the offenders in this study was 23.6 (SD = 6.3), which included four moderate-risk and five high-risk cases based on the standard cut-off criteria constructed by Andrews et al. (2004). Although demographic information for the officers and offenders in these audiotapes were not collected, the officers and offenders in this study consisted of a mixture of males and females. The scores from the four rater assessments for each of the nine offenders (n = 36) and one set of assessments from the two trainers for each of the offenders (n = 9) were then entered into a database for analysis. Given the nature of the study's design, students were unable to access collateral information regarding the offender or file information involving the offender's criminal history that was available to the trainers. We therefore excluded the criminal history questions in this study and focused instead on the remaining seven criminogenic needs domains in our analyses.

Analyses

We conduct several analyses to assess the inter-rater reliability and rater accuracy of the LS/CMI. First, we calculate the average domain deviation scores among the student raters

in each of the criminogenic need subcategories and in the total LS/CMI needs score. The deviation score is the absolute difference between the score for each individual student rater and the mean score for all of the student raters. This value represents the consistency in which the raters scored the offenders similarly in these areas. A deviation score of 0 means that all of the raters scored all of the offenders the same on that construct. This breakdown by subcategory is important because it informs us how consistently each of the specific needs domains are rated. As the number of items vary considerably by domain (range = 2 – 9), we also standardized the deviation scores to make direct comparisons between categories possible. This value is calculated by dividing the average domain deviation score by the number of items within the domain.

Second, our main inter-rater reliability analysis involves the use of the ICC)for the individual domain and the total criminogenic needs scores. The ICC provides a measure of the amount of variance in scores that is due to the variability between raters (Hallgren, 2012). Given that all of the offenders in this study are assessed by the same raters, we use a two-way random-effects model to calculate the ICC values. This model is appropriate for our purposes because it takes into account the variation within raters and within offenders (see Shrout & Fleiss, 1979). We report two types of the ICC, one for "absolute agreement" and one for "consistency." The absolute agreement measure examines the extent to which similar scores are given to the same offender, whereas the consistency measure examines the extent to which raters score offenders in the same direction relative to scores on other offenders. Although both types of ICC are important for assessing inter-rater reliability, we place more emphasis here on absolute agreement. Further, we report the "single measures" version of the ICC rather than "average measures" version because again our interest is to determine how accurate a single rater would be in making ratings on their own, rather than in the average of their ratings across groups. In general, ICC scores range from 0 to 1, with higher scores more indicative of greater reliability. An ICC of 1, for example, indicates that there is no variance between raters on the same offender (i.e., a perfect agreement). Prior research suggests the following guidelines for interpreting the magnitude of the ICC: < .40 = inadequate agreement, .40 to .59 = acceptable agreement, .60 to .74 = good agreement, and .75 to 1.00 = strong agreement (Cicchetti, 1994).[3]

In addition to reliability, it is also important that raters are accurate in their assessment of offender risk. It is possible, for example, that raters could reliably assess offenders, but they may also be consistently inaccurate in their scoring. Accordingly, our third step involves assessing how similar the student rater scores were to the "gold standard" agreed upon trainer scores. We do so by examining the deviation scores and the percent agreement between the student raters and the trainers across each of the need domains. The deviation score is calculated as the absolute difference between the student rater score and the trainer score across each domain. Again, this value is also standardized by item to make direct comparison possible between categories. The percent agreement value represents the extent to which student domain scores were the same as the trainer domain scores. This value is calculated as the number of similar scores between raters and trainers divided by the total number group comparisons. A value of 100% indicates that all of the students have the same total domain score as the trainers. Finally, we examine how reliably and accurately the student raters scored each of the 35 LS/CMI items included in this study. We do so by examining the single rater absolute agreement ICC measure

between the student raters and percent agreement of the student ratings with trainer scores for each item.

Results

We begin by examining the average deviation scores in each of the criminogenic need domains and the total LS/CMI needs score. The deviation values provide information on the variability of scores for each category examined. As can be seen in Table 1,

the average deviation for all offenders' total needs score is 3.14. These scores range across domains, with a low of 0.26 for leisure/recreation to a high of 0.89 for education/employment. It is important to note, however, that the number of items in each domain also vary; leisure/recreation has the fewest items (n = 2) and education/employment has the most (n = 9). Likewise, the magnitude of the deviation scores may not only be influenced by discrepancies between raters, but also by the number of items included within each domain. To address this issue, we standardized the average deviation scores by item to determine how much variation exists in each category per item. This analysis reveals that the domain with the least amount of variation per item is alcohol/drug problem (0.09 per item) and the two domains tied with the most variation per item are leisure/recreation and procriminal attitude (0.13 per item).

We now turn to the ICC analyses, which are also displayed in Table 1. This table shows the absolute agreement and consistency measures. Recall that we are more interested here in the absolute measures because these estimates provide information on the extent to which raters give the same scores to each offender. As can be seen in the table, the total LS/CMI needs score has good reliability in absolute agreement (.64). The ICC for the consistency measure, which measures the extent to which raters score the offenders in the same direction, is also good (.62). The education/employment, leisure/recreation, alcohol/drug, and antisocial pattern domains also display a good level of inter-rater reliability on both of these measures (.62 to .71). The family/marital domain has a strong level of agreement on the absolute (.76) and consistency (.80) measures, and the procriminal attitude domain has an acceptable level of agreement on these two measures (.53 and .54, respectively). Finally, the companions domain has a good level of absolute agreement (.61) and an acceptable level of consistency (.58).

Table 1. Average deviation scores and ICC coefficients for single rater in LS/CMI scores between raters, by domain.

Domain (n of items)	Average Domain Deviation Score	Average Item Deviation Score	Absolute ICC	Consistency ICC
Education/employment (9)	0.89	0.10	.684	.683
Family/marital (4)	0.43	0.11	.761	.804
Leisure/recreation (2)	0.26	0.13	.694	.707
Companions (4)	0.61	0.15	.605	.582
Alcohol/drug problem (8)	0.74	0.09	.635	.630
Procriminal attitude/ orientation (4)	0.51	0.13	.533	.537
Antisocial pattern (4)	0.38	0.10	.668	.649
Total LS/CMI needs score (35)	3.14	0.09	.635	.617

Note. ICC = Intra-class correlation; LS/CMI = Level of Service/Case Management Inventory.

Next, we turn to the rater accuracy analyses. As can be seen in Table 2, the deviation scores between the raters and the trainers range from 0.55 for the companions domain to 1.00 in the procriminal attitude domain. In five of the seven domains—family/marital, leisure/recreation, alcohol/drug, procriminal attitude, and antisocial pattern—the absolute deviation scores between the raters and the trainers is higher in magnitude than the scores found between the four raters (refer to Table 1). This suggests that the raters were more likely to score these domains similarly to one another than they were with the trainers. This is especially pronounced in the domain of procriminal attitude, where the raters have an average deviation score among themselves of .51, and a deviation score of 1.00 with the trainers. As above, we also standardized the deviation scores by item. This analysis reveals that the domain with the most similar ratings between students and trainers per item was education/employment (0.08 deviation per item) and the least similar was leisure/recreation (0.29 deviation per item).

Table 2 also shows the percent agreement in the domain scores between the raters and the trainers. As can be seen in the table, three of the domains have approximately a 30% agreement: family/marital, procriminal attitude, and antisocial pattern. Interestingly, procriminal attitude and family/marital have the lowest percent agreement scores, but these two domains also have the lowest and highest ICC scores (refer to Table 1). As stated above, the number of items vary across domains, which may have some bearing on these results. One may anticipate, for example, that there would be less of a percent agreement in the domains with a greater range of possible scores (e.g., education/employment [range of 0 – 9 points] versus leisure/recreation [range of 0 – 2 points]).

Finally, we examine the reliability and accuracy across the 35 LS/CMI criminogenic need items. Table 3 reports the more conservative ICC value (i.e., absolute ICC) for the student ratings, and the percent agreement between the raters and trainers by item. There were 21 items that had a good or strong level of agreement (i.e., ICC > .60) and seven items with an inadequate level of consistency (i.e., ICC < .40). Further, 13 of the items were scored by raters with more than a 70% agreement with the trainers, and six that had less than 50% agreement with the trainers' scores.

This individual-item level analysis also affords the ability to assess which specific questions may be attributable for the low reliability within the broader domain categories. For example, the procriminal attitude domain was found to have the lowest absolute and consistency ICC values (refer to Table 1). However, inspection of Table 3 shows that the issues of reliability stem primarily from Question 39 and to a lesser extent Question 37. This also appears to be the case in the companions domain, which has the next lowest set of ICC values, where the low performance in domain reliability appears to be largely driven by the low reliability in questions 26 and 25. A similar pattern of findings emerges

Table 2. Average deviation scores and percent agreement in Level of Service/Case Management Inventory scores between raters and trainers, by domain.

Domain	Average Domain Deviation Score	Average Item Deviation Score	% Agreement
Education/employment	0.75	0.08	44.4
Family/marital	0.84	0.21	30.6
Leisure/recreation	0.57	0.29	43.8
Companions	0.55	0.14	42.9
Alcohol/drug problem	0.91	0.11	41.7
Procriminal attitude/orientation	1.00	0.25	30.6
Antisocial pattern	0.61	0.15	31.3

Table 3. ICC coefficients for single rater in LS/CMI Scores between raters, and percent agreement in scores between raters and trainers, by item.

Item	Absolute ICC	% Agreement
Education/employment		
9. Currently unemployed	.613	75.0
10. Frequently unemployed	.754	50.0
11. Never employed a full year	1.000	56.3
12. Less than grade 10 or equivalent	1.000	100.0
13. Less than grade 12 or equivalent	1.000	100.0
14. Suspended or expelled at least once	1.000	87.5
15. Participation/performance	.579	78.1
16. Peer interactions	.725	78.1
17. Authority interaction	.725	78.1
Family/marital		
18. Dissatisfaction with marital or equivalent situation	.863	53.1
19. Nonrewarding parental	.355	56.3
20. Nonrewarding, other relatives	.493	43.8
21. Criminal family/spouse	1.000	100.0
Leisure/recreation		
22. Absence of recent participation in organized activity	.695	68.8
23. Could make better use of time	.559	18.8
Companions		
24. Some criminal acquaintances	.736	78.1
25. Some criminal friends	.306	46.9
26. Few anticriminal acquaintances	.067	46.9
27. Few anticriminal friends	.551	43.8
Alcohol/drug problem		
28. Alcohol problem, ever	.679	87.5
29. Drug problem, ever	1.000[a]	100.0
30. Alcohol problem, currently	.857	71.9
31. Drug problem, currently	.613	34.4
32. Law violations	.667	65.6
33. Marital/family	.538	53.1
34. School/work	.598	65.6
35. Medical or other clinical indicators	.111	59.4
Procriminal attitude/orientation		
36. Supportive of crime	.818	56.3
37. Unfavorable toward convention	.307	43.8
38. Poor, toward sentence/offense	.702	53.1
39. Poor, toward supervision/treatment	.141	68.8
Antisocial pattern		
40. Specialized assessment for antisocial pattern	1.000[a]	50.0
41. Early and diverse antisocial behavior	.892	84.4
42. Criminal attitude	.377	53.1
43. Pattern of generalized trouble	.514	28.1

Note. ICC = Intra-class correlation. LS/CMI = Level of Service/Case Management Inventory.
[a]Raters scored these questions as constants.

with respect to the percent agreement analyses. For example, the two categories with the lowest percent agreement in the total domain score between student raters and trainers are family/marital and procriminal attitude (refer to Table 1); however, this low level of domain score agreement appears heavily influenced by the low level of percent agreement in Questions 20 and 37, respectively.

Finally, the ICC values for seven of the 35 items was found to be 1.00. Although this result indicates perfect agreement on 20% of the items, it should be noted that two of these

items were scored as constants. More specifically, the students rated all of the offenders as having a drug problem at some point in their lives (Question 29) and none of the offenders as having evidence of an antisocial personality disorder (Question 40). This lack of variability found in these two questions may indicate a problem with these items; however, it may also reflect a lack of heterogeneity among the nine offenders in this study.

Limitations

Although the current study has made several advances, it nonetheless has limitations that should be understood and addressed in future research. For instance, the raters involved in this study included university students trained alongside probation officers, not correctional professionals with years of experience in the field. It is important to note that several steps were taken to ensure that the students were trained in the same manner as officers, and anecdotally the trainers reported that the students seemed to be as knowledgeable in scoring as the officers who typically go through the training. This is particularly pertinent to note because newly hired officers are likely in a similar standing as the students in terms of experience. Thus, though not necessarily generalizable to long-experienced officers, the findings are perhaps relatable to the use of LS/CMI interviews among new officers.

Second, this study used audio-recordings of officers conducting LS/CMI interviews with offenders. This strategy was helpful in reducing the potential for a test–retest bias because it meant that the same person did not need to be interviewed multiple times; however, it also fails to address the fact that different raters may elicit information in different ways from offenders. Such individual strategies may allow for an interviewing officer to ask certain questions to better contextualize the offender's responses in a way that increases rater accuracy. Future research should seek to further unpack whether differences in interview strategies and styles result in the collection of different types of information from offenders that would result in differences in scoring.

Finally, our measure of accuracy relied on the scores provided by the trainers in this study. Although one may argue that such absolute information in unknowable, we maintain that because trainers have spent the most amount of time reviewing the scoring manuals, it is their opinions that should be viewed as the gold standard. That being said, another future study might test this assumption by examining the inter-rater reliability in a group of trainers. We hypothesize that such a test would yield much higher estimates of reliability while also providing an average collective benchmark to which accuracy assessments among officers can be gauged.

Discussion

Ensuring a high level of inter-rater reliability is important in the area of offender risk assessment because risk scores are used to determine a multitude of discretionary legal dimensions, including sentencing, supervision, and treatment conditions. The purpose of the current study was to assess the inter-rater reliability and rater accuracy of the LS/CMI. This work sought to address some of the limitations from the previous inter-rater reliability studies in this area. The design of this study is important because it included multiple raters and multiple offenders and used measures of percent agreement and ICC at the domain and item level. The results of the study revealed an adequate to strong level

of inter-rater reliability across the criminogenic need domains of the LS/CMI. These results also suggest there is a wide range of rater accuracy across the domains and items of the LS/CMI.

Despite the limitations noted above, we believe that this study has much to offer in terms of practical recommendations and points of consideration. To obtain a high level of inter-rater reliability and accuracy, for example, raters must be trained well on the tool and the items must be clearly defined. One implication of these findings is that trainers may choose to spend more time in the initial training going over the items found to have lower inter-rater reliability and accuracy. Further, it may also be beneficial for trainers to make these items the topic of subsequent yearly booster training sessions. Another implication is that agencies add clarity to the guidelines of the confusing items to be clearer about the scoring criteria. It may also be worthwhile for agencies to develop a semistructured interview process with must ask questions and enough flexibility to ensure that sufficient information is collected to accurately and reliably score the LS/CMI.

Although the findings of this study do not offer a solution as to how to improve officer interviewing strategies, they do shed an important light on which criminogenic need domains and items have the lowest and highest levels of reliability. For example, compare the findings related to the domains of procriminal attitudes and companions to that of family martial and leisure/recreation. Although they largely fell into the "adequate" range, the domains of procriminal attitudes and companions yielded the lowest absolute and consistency ICC measures (see Table 1). This is understandable considering these domains aim to capture dynamic and abstract concepts that are difficult to measure. In contrast, family/marital and leisure/recreation yielded the highest values in the same ICC measures. This suggests that these domains may be more discrete concepts and easier to measure, and/or they are being captured in a more structured manner within the administration of the assessment. These distinctions are most apparent in Table 3, where the absolute ICC is provided for each item. Those questions with the some of the lowest ICC values as well as lowest percent agreement with the trainers fell into the items that are qualitatively more difficult to measure.

Regardless as to why the reliability of domains may vary, it is important to note that none of the subcategories as a whole had an ICC value that would suggest inadequate reliability. The same can not be said for the individual items, however. Setting aside the potential debate of whether "adequate" reliability of a domain or item within an offender risk/needs assessment is an issue of concern, there were a seven individual items that fell into the area of "inadequate" reliability (i.e., ICC < .40; see Table 3). These items all appear to highlight important yet abstract or dynamic areas that are difficult to measure. Such findings indicate that these items, or the training surrounding these items, should be strengthened to increase reliability in some fashion. This finding also suggests that the instrument appears to have more of an issue with reliability at the item level and not so much at the domain level.

Finally, the results of this study appear to support the findings of other literature. Compared to the findings of Austin et al. (2003) for example, it appears that the deviation of three points in the total score may be rather consistent. It should be kept in mind however, that Austin et al. used a different version of the assessment (LSI-R), included criminal history (nine additional items), and only had two raters. Our interpretation of those three point differences in the total score however, is not necessarily one of failure on the side of the

assessment. As suggested at the beginning of this article, where the three-point difference exists along the cut-points of an instrument (i.e., low, moderate, or high) will determine if it is very meaningful. It is possible that the three points could mean the difference between high and moderate supervision contacts as well as being referred to treatment or not. It is just as possible that the three-point difference falls within a category of risk. That is, the difference of having a 44 or 41 in the total score is minimal because they may be in the high-risk category. In this sense, it depends on how and where the cut-points are made, but this is an area for future research and one that is outside the scope of this article.

In conclusion, we believe that this work provides much needed attention in the neglected research area of reliability in offender risk assessment. This study falls on the heels of the recent works by Michael Rocque, Plummer-Beale, and Grant Duwe and stresses to correctional researchers, administrators, and practitioners that reliability is an important aspect in the selection and use of an offender risk assessment. It is our hope that more attention will be paid to reliability in the future and that conversations regarding which assessment to choose will include talks of reliability measures along with predictive accuracy measures.

Notes

1. For more on the discussion of the generations in risk assessments see Bonta and Andrews (2017, pp. 192–222).
2. Most of the other scholarship on reliability of the LS instruments are found in unpublished conference presentations, theses, and dissertations. These other works primarily involve examinations of alpha (α) levels, which can help inform how consistently the items within a domain are scored as a group, but not how reliably the assessment is scored by raters. For more information on this research see Andrews et al. (2010).
3. Some researchers suggest a different criterion for interpreting the ICC values. Baird et al. (2013), for example, use 0.0 to 0.2 = *poor*, 0.3 to 0.4 = *fair*, 0.5 to 0.6 = *moderate*, 0.7 to 0.8 = *strong*, and > 0.8 = *almost perfect*.

Acknowledgement

The authors wish to thank Kimberly Bernard for her editorial comments. The authors also wish to thank the two anonymous reviewers for their helpful and constructive comments.

References

Andrews, D. A. (1982). *The Level of Service Inventory (LSI): The first follow-up*. Toronto, Canada: Ministry of Correctional Services.

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

Andrews, D. A., Bonta, J., & Hoge, R. D. (1990). Classification for effective rehabilitation: Rediscovering psychology. *Criminal Justice and Behavior*, *17*(1), 19–52.

Andrews, D. A., Bonta, J., & Wormith, J. S. (2004). *The Level of Service/Case Management Inventory (LS/CMI)*. Toronto, Canada: Multi-Health Systems.

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

Andrews, D. A., Bonta, J., & Wormith, J. S. (2010). The Level of Service (LS) assessment of adults and older adolescents. In R. K. Otto & K. S. Douglas (Eds.), *Handbook of violence risk assessment* (pp. 199–226). New York, NY: Routledge.

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

Andrews, D. A., & Dowden, C. (2007). The risk-need-responsivity model of assessment and human service in prevention and corrections: Crime prevention jurisprudence. *Canadian Journal of Criminology and Criminal Justice*, *49*, 439–464.

Andrews, D. A., Guzzo, L., Raynor, P., Rowe, R. C., Rettinger, L. J., Brews, A., & Wormith, J. S. (2012). Are the major risk risk/needs factors predictive of both female and male reoffending? A test with the eight domains of the Level of Service/Case Management Inventory. *International Journal of Offender Therapy and Comparative Criminology*, *56*(1), 113–133.

Andrews, D. A., Zinger, I., Hoge, R. D., Bonta, J., Gendreau, P., & Cullen, F. T. (1990). Does correctional treatment work? A clinically-relevant and psychologically-informed meta-analysis. *Criminology*, *28*(3), 369–404.

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

Austin, J., Coleman, D., Peyton, J., & Johnson, K. D. (2003). *Reliability and validity study of the LSI-R risk assessment instrument*. Washington, DC: Institute of Crime, Justice and Corrections.

Baird, C. (2009). *A question of the evidence: A critique of risk assessment models used in the justice system*. Madison, WI: National Council on Crime and Delinquency.

Baird, C., Healy, T., Johnson, K., Bogie, A., Dankert, E. W., & Scharenbroch, C. (2013). *A comparison of risk assessment instruments in juvenile justice*. Washington, DC: Office of Juvenile Justice and Delinquency Prevention.

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

Bonta, J., & Andrews, D. A. (2017). *The psychology of criminal conduct* (6th ed.). New York, NY: Routledge.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychology Assessment*, *6*(4), 284–290.

Cohen, T. H. (2017). Automating risk assessment instruments and reliability: Examining an important but neglected area in risk assessment research. *Criminology and Public Policy*, *16*(1), 235–269.

Cullen, F. T., & Gendreau, P. (2001). From nothing works to what works: Changing professional ideology in the 21st century. *The Prison Journal*, *81*(3), 313–338.

Desmarais, S. L., & Singh, J. P. (2013). *Risk assessment instruments validated and implemented in correctional settings in the United States*. Bethesda, MD: Justice Center, The Council of State Governments.

Duwe, G., & Rocque, M. (2017). Effects of automating recidivism risk assessment on reliability, predictive validity, and return on investment (ROI). *Criminology and Public Policy*, *16*(1), 271–279.

Gendreau, P. (1996). The principles of effective intervention with offenders. In A. T. Harland (Ed.), *Choosing correctional options that work: Defining the demand and evaluating the supply* (pp. 117–130). Thousand Oaks, CA: Sage.

Girard, L., & Wormith, J. S. (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.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23–34.

Hoge, R. D., & Andrews, D. A. (2002). *The Youth Level of Service/Case Management Inventory*. Toronto, Canada: Multi-Health Systems.

Hollin, C. R. (2002). Risk-needs assessment and allocation to offender programmes. In J. McGuire (Ed.), *Offender rehabilitation and treatment: Effective programmes to reduce re-offending* (pp. 309–332). Chichester, UK: Wiley.

Latessa, E. J., & Lovins, B. (2010). The role of offender risk assessment: A policy maker guide. *Victims and Offenders*, *5*(3), 203–219.

Lipsey, M. W., & Cullen, F. T. (2007). The effectiveness of correctional rehabilitation: A review of systematic reviews. *Annual Review of Law and Social Science*, *3*, 297–320.

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., & Holsinger, A. M. (2006). The risk principle in action: What have we learned from 13,676 offenders and 97 correctional programs? *Crime and Delinquency*, *52*(1), 77–93.

Olver, M. E., Stockdale, K. C., & Wormith, J. S. (2014). Thirty years of research on the Level of Service scales: A meta-analytic examination of predictive accuracy and sources of variability. *Psychological Assessment*, *26*(1), 156–176.

Rocque, M., & Plummer-Beale, J. (2014). In the eye of the beholder? An examination of the inter-rater reliability of the LSI-R and YLS/CMI in a correctional agency. *Journal of Criminal Justice*, *42*(6), 568–578.

Schmidt, F., Hoge, R. D., & Gomes, L. (2005). Reliability and validity analyses of the Youth Level of Service/Case Management Inventory. *Criminal Justice and Behavior*, *32*(3), 329–344.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428.

Singh, J. P., Fazel, S., Gueorguieva, R., & Buchanan, A. (2014). Rates of violence in patients classified as high risk by structured risk assessment instruments. *British Journal of Psychiatry*, *204*, 180–187.

Smith, P. (2013). The psychology of criminal conduct. In F. T. Cullen & P. Wilcox (Eds.), *The Oxford handbook of criminological theory* (pp. 69–88). New York, NY: Oxford University Press.

Taxman, F. S. (Ed.). (2017). *Handbook on risk and need assessment: Theory and practice*. New York, NY: Routledge.

Wormith, J. S. (2011). The legacy of D. A. Andrews in the field of criminal justice: How theory and research can change policy and practice. *International Journal of Forensic Mental Health*, *10*(2), 78–82.

Wormith, J. S., Hogg, S. M., & Guzzo, L. (2012). The predictive validity of a general risk/needs assessment inventory on sexual offender recidivism and an exploration of the professional override. *Criminal Justice and Behavior*, *39*(12), 1511–1538.

Wormith, J. S., Hogg, S. M., & Guzzo, L. (2015). The predictive validity of the LS/CMI with Aboriginal offenders in Canada. *Criminal Justice and Behavior*, *42*(5), 481–508.