The current study aimed to provide data on the performance of items, dimensions, and the total score of the Level of Service/Case Management Inventory (LS/CMI), one of the most internationally used actuarial scales for the prediction of general recidivism in convicted persons. Using the full population of Quebec’s male incarcerated population evaluated between 2008 and 2015 with a 2-year follow-up \((N = 15,961)\), results indicated that the predictive validity of the scale and its components was in line or better than effect sizes reported in other validation studies. A Rasch model was computed to obtain the difficulty parameter of LS/CMI items. Results indicated that items had varying levels of difficulty and covered the whole spectrum of the risk continuum. However, difficulty in Rasch was uncorrelated with the predictive validity of items, which casts a doubt on the applicability of some aspects of item response theory to actuarial scales.

**Keywords:** Level of Service/Case Management Inventory (LS/CMI); predictive validity; Rasch; item response theory; actuarial scales; general recidivism; convicted persons; incarcerated persons

Risk assessment for crime and violence has grown at a very fast pace during the last 30 years, especially in North America. Following the media coverage of particularly sordid sexual aggressions and murders by recidivists, decision-makers from Canadian and American governments funded the development and implementation of reliable and valid risk assessment procedures and correctional interventions based on the risk-need-responsivity model (Andrews & Bonta, 2010; Brouillette-Alarie & Lussier, 2018).

The first wave of reliable risk assessment tools is known as static actuarial assessment, a method that succeeded unstructured clinical judgment with its lackluster inter-rater agreement and poor predictive validity (Dawes et al., 1989; Grove et al., 2000). Actuarial

**AUTHORS’ NOTE:** The views expressed are those of the authors and not necessarily those of the Ministère de la Sécurité publique du Québec. Correspondence concerning this article should be addressed to Guy Giguère, Ministère de la Sécurité publique du Québec, 2525 Boulevard Laurier, Quebec City, Quebec, Canada G1V 2L2; e-mail: guy.giguere@msp.gouv.qc.ca.
assessment reliably determines the level of risk by mechanically combining empirically validated predictors. This method is considered “atheoretical,” because the main inclusion criterion of an item in a scale is its statistical association with the outcome of interest, not its theoretical relevance. The first actuarial scales comprised static risk factors only, which strengthened the perception that these tools were largely atheoretical (Andrews & Bonta, 2010; Bonta, 1996). The second wave of actuarial scales was defined by its inclusion of dynamic risk factors, also known as criminogenic needs. Therefore, they were better positioned to follow the evolution of risk over time and suggest intervention targets than instruments that exclusively comprised static risk factors (Andrews & Bonta, 2010; Bonta, 1996, 2002; Gendreau et al., 1996; Hanson & Harris, 2000). The most recent generation of actuarial scales is case management risk/need tools (Andrews & Bonta, 2010). In addition to assessing static and dynamic risk factors, this generation provides clear guidelines to ensure that case management will be consistent with the results of risk assessment, according to the risk and need principles of effective correctional intervention. The Level of Service/Case Management Inventory (LS/CMI; Andrews et al., 2004) is a prime example of a case management risk/need tool that assesses general recidivism risk.

Having reliable and valid risk assessments has numerous advantages. Overestimation of risk can lead to the long-term imprisonment of individuals who could otherwise become productive members of society or actively impair their chances of reintegration upon release. Indeed, high-risk statuses, such as “sexually violent predator” are known to be significant obstacles to reintegration, limiting housing, and employment opportunities (Harris, 2014). Conversely, underestimation of risk can lead to the release of dangerous individuals and result in new victims. Therefore, precise risk assessments that are neither too high nor too low have been increasingly seen as a cornerstone of correctional practice since the early 1990s (Brouillette-Alarie & Lussier, 2018).

THE FOCUS OF RISK TOOLS ON PREDICTIVE VALIDITY AND THE NEGLECT OF OTHER PSYCHOMETRIC ASPECTS

Because the primary objective of actuarial scales is to make the most accurate predictions to inform risk management and intervention efforts, studies on risk tools have traditionally focused on predictive validity rather than construct validity or other psychometric aspects (Helmus & Babchishin, 2017). Contrarily to psychometric tests from the field of psychology or ability tests from the field of education, criminological risk tools are more interested in making an accurate prediction about the risk of reoffending than determining the ability of an individual on a construct (e.g., extraversion/introversion, algebra skills). Therefore, risk tool validation studies have typically neglected relevant sources of evidence that are potentially necessary to justify the interpretation and uses of scores stemming from actuarial scales (Helmus & Babchishin, 2017; Messick, 1989). At the forefront of these sources of evidence lies construct validity, the degree to which a test measures what it claims to be measuring (Cronbach & Meehl, 1955).

Many authors have advocated for the integration of construct-oriented approaches in criminological risk assessment practice (Babchishin et al., 2016; Brouillette-Alarie et al., 2016, 2022; Mann et al., 2010). Clarifying the construct validity of risk tools has many potential advantages for the field. First, it offers insight into why certain scales predict certain outcomes better than others, as this is dependent on the constructs they assess and how
each construct is weighted in these scales. This, in turn, can help evaluators integrate the potentially conflicting results of risk scales when multiple tools are available for the same population but arrive at different conclusions (Barbaree et al., 2009). Second, understanding the constructs implicit in risk tools can improve predictive accuracy. Specifically, when the constructs are known, it is possible to improve the reliability and validity of their assessment using standard psychometric methods and, therefore, improve the predictive accuracy of scales (Brouillette-Alarie et al., 2016). Finally, construct-level approaches maximize the clinical relevance of existing scales by focusing on psychological dimensions and their nomological network, facilitating the identification of the “source” of the risk. Evaluations that address psychological features are generally better received by clinicians, practitioners, and decision-makers than those that only delineate the level of risk (Mann et al., 2010).

Over the last 10 to 15 years, studies have increasingly considered the construct validity of risk tools by using standard psychometric methods, such as factor analysis and convergent/discriminant validity analyses (Babchishin & Hanson, 2020; Brouillette-Alarie et al., 2016, 2018; Brouillette-Alarie & Proulx, 2019; Gordon et al., 2015). For example, the latent constructs of the Static-99R and Static-2002R (Hanson & Thornton, 2000, 2003; Helmus et al., 2012)—static risk tools for persons convicted of a sexual offense—were studied in a research program that led to a three-factor model comprising sexual criminality, general criminality, and “youthful stranger aggression” (a factor comprising items related to youth and victim harm; Brouillette-Alarie et al., 2016). Then, the nomological network of these static dimensions was studied by linking them with psychologically relevant items/constructs and recidivism outcomes (Brouillette-Alarie et al., 2018; Brouillette-Alarie & Proulx, 2019).

General criminality risk tools, such as the LS/CMI have also been subjected to factor analyses, though less often than risk tools for individuals convicted of a sexual offense. Gordon et al. (2015) reported that there were few available factor analyses of the LS/CMI and referred readers to analyses of the Level of Service Inventory—Revised (LSI-R; Andrews & Bonta, 1995), its predecessor, for insight into the latent structure of risk tools for “general” convicted persons. Studies of the LSI-R’s factor structure are not numerous and have been criticized due to their methodological choices (e.g., Andrews & Robinson, 1984; Arens et al., 1996). Indeed, these studies mostly used principal component analysis over factor analysis and entered dimensions rather than items in the analysis, a surprising choice that seemed to be driven by the desire to obtain a single resulting factor (Hsu et al., 2011). To correct the aforementioned limitations, Hsu et al. (2011) conducted a factor analysis of LSI-R items and obtained five dimensions for men and four for women. The four dimensions common to both genders were static risk, employment, pro-criminal attitudes, and mental health. The fifth dimension, exclusively present in men, was protective companions.

THE RASCH MODEL AND ITS POTENTIAL RELEVANCE FOR RISK TOOL VALIDATION

Another important and relatively recent means to study construct validity has been item response theory (IRT) and Rasch models. Scholars from the field of education have long advocated the use of such models, as they are less sensitive than classical test theory models to circular dependency (i.e., dependence on the overall performance of the validation sample; de Ayala, 2009). IRT was introduced in the 1950s and 1960s (Lord, 1953; Rasch, 1960)
the LS/cMi

The LS/CMI is the evolution of the Level of Service Inventory-Revised (LSI-R) and is the most heavily used actuarial scale for the prediction of general recidivism internationally (Wormith, 2011). It relies on the substantial body of work by Don Andrews and James Bonta and their theory of the psychology of criminal conduct (Andrews & Bonta, 2010). The scale has been validated with men, women, adults, adolescents, incarcerated persons, and persons on parole (for reviews, see Andrews & Bonta, 2010; Olver et al., 2014). It is to better assess item difficulty and discrimination and create sample-free measures (Osteen, 2010). The Rasch model can be seen as a continuous line on which individuals and items are both placed according to their respective skill level/difficulty. Applied to the field of criminology, the Rasch model could establish which items are more difficult to endorse for individuals whose “disposition toward crime” is under study; endorsing an item considered difficult could enhance an individual’s disposition to crime more than an item everyone is likely to endorse. When analyzing data through the Rasch model, fit statistics are calculated, showing which items are too predictable or too unstable to help the model fit the data. Misfitting items can then be kept, discarded, or reviewed by experts (Bohlig et al., 1998).

Therefore, IRT and Rasch models could offer interesting avenues for improving existing risk scales or making shorter versions of them by removing unfitting or redundant items. Even though complete assessments of risk and needs are generally preferable to screening versions, short versions of risk scales can be relevant for jurisdictions where time and resources are limited. Screening versions of the LSI-R (Andrews & Bonta, 1998) and Psychopathy Checklist-Revised (Hart et al., 1995) have been developed, and IRT techniques could help to review or contextualize decisions that were made in creating short versions of these scales.

Another potential advantage of IRT-based models concerns refined item weighting. As of now, most actuarial scales comprise items worth one point each that are summed to determine the total risk score. However, as it pertains to face validity, it is unlikely that items on actuarial scales are all equally difficult and, thus, equally risk relevant. Therefore, it is not impossible that a more refined weighting of items according to their difficulty could enable improvements in predictive validity (Brouillette-Alarie et al., 2022). There are debates on the tangible benefits of differentially weighting items, with some results indicating that complex combinations rarely outperform the simple summing of dichotomous items (Ghiselli et al., 1981; Grann & Långström, 2007; Silver et al., 2000). However, differential weighting has its greatest impact when there is a wide variation in weighting values, little intercorrelation between items, and only a few items (Ghiselli et al., 1981; Kline, 2005). Considering that actuarial scales usually comprise few nonredundant items, they could potentially benefit from differential weighting (Georgiou, 2019).

Another overarching benefit of Rasch and IRT models is that they bring the focus to test items rather than total scores. Even though classical test theory comprises techniques to assess the performance of individual items, for example, difficulty calculations and item-total correlations, these tests tend to be underreported in non-IRT articles in the criminological literature. As the next section will illustrate, some of the most heavily used actuarial scales in corrections report very few (if any) data on the performance of their individual items.
one of the few fourth-generation risk tools, as it integrates case management elements in addition to third-generation risk assessment procedures (Brouillette-Alarie & Lussier, 2018). Case management sections are based on the tried-and-true risk, need, and responsivity principles of effective correctional intervention (Andrews & Bonta, 2010).

The LS/CMI comprises eight risk domains that have all demonstrated predictive validity towards general recidivism (Andrews & Bonta, 2010; Andrews et al., 2011; Olver et al., 2014). In the Canadian context, the predictive validity of its total score rivals that of the best tools in the field, with area under the curve (AUCs) reaching .75 or more depending on the validation sample (Andrews et al., 2011; Olver et al., 2014). In samples from the United States, the predictive validity of the LS measures in general was found to be much lower (Olver et al., 2014). Despite the substantial literature on the dimensions and total scores of the LS/CMI, very few (if any) data are available on the performance of its individual items. We are aware of no studies that relate to the predictive validity of LS/CMI items and only two studies that report other parameters, such as item-total correlations, difficulty, and discrimination (Giguère et al., 2015; Giguère & Lussier, 2016). The first study looked at these parameters using classical test theory, and the second one used two-parameter IRT. In the latter, Giguère and Lussier (2016) found that many LS/CMI items were redundant or displayed problematic discrimination and/or difficulty values. After removing the problematic items, they found that the remaining items achieved a predictive validity that was very close to that of the full 43 items.

Another application of IRT for risk tool validation can be found in Huang et al. (2021), who investigated the generalizability of the Youth Level of Service/Case Management Inventory (YLS/CMI; Hoge & Andrews, 2011) using a sample of Indigenous and non-Indigenous youth. Differential item functioning analyses indicated that items from the education domain were less likely to be endorsed by Indigenous youth, while items from the substance abuse domain were more likely to be endorsed. Importantly, predictive validity analyses revealed that the YLS/CMI was not predictive of criminal recidivism for Indigenous youth.

Even though we commend the LS/CMI for its sound theoretical underpinnings and substantial validation as it relates to its dimensional and total scores, the lack of publicly available data on the performance of its items is a significant limitation that needs addressing. Data on the psychometric properties of individual LS/CMI items could enable the identification of problematic items, which could in turn lead to item deletion, reworking, or reweighting, and, hopefully, improvements in predictive validity. It could also lead the way for the development of a newer and shorter version of the LS/CMI, if problematic items were to be found.

OBJECTIVES

The current study aimed to address the lack of validation data on individual LS/CMI items using the Rasch model and predictive validity analyses. The first step of our examination was to enter the 43 LS/CMI items in the Rasch model and see which ones did not fit the model (or latent trait). Second, the difficulty parameter of each remaining item was computed. Third, a Wright map of item difficulty and person ability (Wright & Stone, 1979) was drawn to better visualize Rasch results. Fourth, the predictive validity of LS/CMI items, dimensions, and total score was tested in relation to recidivism with a 2-year follow-up.
Finally, the predictive validity of items was correlated to their difficulty to verify whether difficulty in Rasch equates to disposition toward crime (predictive validity toward recidivism). Results of our analyses will be used to discuss potential improvement pathways for the LS/CMI and the relevance of IRT-based models for risk tool validation.

METHOD

SAMPLE AND DATA COLLECTION

In 2007, when the Act Respecting the Quebec (Canada) Correctional System came into effect, a computerized system was established to allow probation officers and prison counselors to compile information from the completed LS/CMIs of convicted individuals. For this study, the sample was taken from the Évaluation des risques et des besoins (ERB) database of Quebec’s Department of Public Safety. This computerized management system enables correctional service staff to easily access convicted persons’ files for court lighting activities or correctional intervention planning.

In the ERB database, the same individual could be found in multiple entries, as each of their contacts with the criminal justice system was entered in one row. To ensure that each individual would be counted once in the analyses, we kept only the most recent record of each convicted individual. The final sample ($N = 15,961$; mean age $= 37.13\ [SD = 12.28]$) ended up depicting the whole population of Quebec’s incarcerated men registered and evaluated between March 2008 and October 2015. Therefore, our sample can be considered representative of Quebec’s recent practices in correctional risk assessment. Convicted individuals in our dataset all received a sentence of less than 2 years for a criminal offense, which means that they were under the supervision of Quebec’s Department of Public Safety. The ERB database did not comprise data on the race and ethnicity of participants. Even though this constitutes a limitation, the majority of participants can be assumed to be White. There were no missing data on the variables used in the statistical analyses.

Because LS/CMI norms are different for incarcerated persons and individuals under parole, we chose, for length and clarity purposes, to present data exclusively on incarcerated persons. Merging these groups together would have been at odds with official LS/CMI documentation (Andrews et al., 2004), and reporting results for both groups would have made this study excessively long, as a separate Rasch model would have been necessary for each group. Obtaining item-level data on individuals under parole is nevertheless an important endeavor that needs to be undertaken in the future.

MEASURES

The French Version of the LS/CMI

The Level of Service/Case Management Inventory (LS/CMI; Andrews et al., 2004) is an assessment and case management tool that measures the risk and need factors of late adolescent and adult convicted persons. The section of the LS/CMI relevant to our investigation is the “General Risk/Need Factors.” This section contains 43 items sorted under the following dimensions: Criminal History (8 items); Education/Employment (6 items); Family/Marital (4 items); Leisure/Recreation (2 items); Alcohol/Drug Problem (8 items); Procriminal Attitude/Orientation (4 items); and Antisocial Pattern (4 items). Each item is coded on a binary response scale (present or absent) by a probation officer or prison
counselor who conducts an interview with the person and consults their criminal record. The total score thus ranges from 0 to 43 points. The total risk and dimensional scores can be used to guide surveillance, determine release conditions, plan and deliver appropriate interventions, and modulate intervention intensity. The French version of the LS/CMI was developed using a cross-cultural procedure. The translated version was translated back into English, and both versions were submitted to the developers of the LS/CMI for approval (Guay, 2016).

**Criminal Recidivism**

In this study, recidivism was considered to occur when an individual who had been previously convicted committed a new crime upon release. The follow-up period was of 2 years, implying that data were right censored (i.e., if recidivism happened after 2 years of follow-up, it would count as no recidivism in the analyses). Breach of conditions was not considered as a new conviction. In our sample, 29.8% of men who had been sentenced to detention reoffended in the 2-year follow-up period. In conformance with previous studies that examined the predictive value of risk evaluation instruments, only time at risk for criminal recidivism was considered (see Giguère & Lussier, 2016). This implies that the follow-up period began as soon as individuals were released from the detention center. Descriptive statistics of our study can be found in Table 1.

**ANALYTICAL STRATEGY**

**Rasch Model**

The classical Rasch model (Rasch, 1960) is a unidimensional measurement model that mathematically represents the relationship between item difficulty and a person’s ability to allow predictions based on the difference in logits between the two:

\[
P(\theta) = \exp(\theta - b) / (1 + \exp(\theta - b))
\]

The basic principle is that one’s probability to succeed an item will be higher if their ability (\(\theta\)) exceeds the item’s difficulty (\(b\)), and it will be lower if the item’s difficulty is greater. Since these two parameters are independent of each other, they give measures invariance, which allows the use of the same item parameters with other comparable samples of individuals—a property that classical test theory models do not support (Boone, 2016; Engelhard & Wang, 2021; Iramaneerat et al., 2008).

The Rasch model produces an asymptotic sigmoid curve named the item characteristic curve (ICC) that represents the probability of succeeding to an item of a given difficulty at different ability levels (\(\theta\)). While other IRT models take into account other parameters, such as discrimination (\(a\)) or pseudo-guessing (\(c\)), the Rasch model considers those as noise and sets them at constant values. The Rasch model was preferred to the two-parameter IRT because it does not comprise any assumptions about the distribution of the latent trait in the population. Two-parameter IRT assumes a normal distribution of the latent variable, which can be unfitting with criminological data, as such data are often positively skewed (Osgood et al., 2002). Using the Rasch model (or any IRT model) comes with a few assumptions that need to be checked to have a good model-data fit.
| Variables | M (SD)% |
|-----------|---------|
| Age       | 37.13 (12.28) |
| Recidivism| 29.8%   |
| LS/CMIa   |         |
| CH1: Any prior conviction | 88.2% |
| CH2: Two prior convictions  | 80.3% |
| CH3: Three prior convictions | 73.1% |
| CH4: Three or more offenses for the current sentence | 72.0% |
| CH5: Arrested/charged before being 16 years old | 29.1% |
| CH6: Ever incarcerated | 92.7% |
| CH7: Institutional misconduct | 46.5% |
| CH8: Breach of probation/parole conditions | 75.3% |
| Criminal History total | 5.57 (2.03) |
| EE9: Currently without a job | 66.6% |
| EE10: Frequently without a job | 54.2% |
| EE11: Never maintained a job for 1 year | 25.4% |
| EE12: Did not achieve grade 10 | 40.6% |
| EE13: Did not achieve grade 12 | 72.5% |
| EE14: Suspended/expelled from school | 48.0% |
| EE15: Not motivated/successful at school/work | 71.4% |
| EE16: Problematic relationships with peers at school/work | 68.8% |
| EE17: Problematic relationships with authority at school/work | 69.1% |
| Education/Employment total | 5.17 (2.75) |
| FM18: Not satisfied with intimate relationship or lack thereof | 33.6% |
| FM19: Problematic relationship with parents | 59.1% |
| FM20: Problematic relationship with other relatives | 44.7% |
| FM21: A family member or spouse has a criminal record | 43.5% |
| Family/Marital total | 1.81 (1.16) |
| LR22: No prosocial activities | 84.3% |
| LR23: Has too much free time | 67.3% |
| Leisure/Recreation total | 1.52 (.66) |
| CO24: Links with criminalized individuals | 99.9% |
| CO25: Friends with criminalized individuals | 44.9% |
| CO26: Few prosocial links | 43.0% |
| CO27: Few prosocial friends | 73.4% |
| Companions total | 2.61 (1.02) |
| ADP28: Ever had problems with alcohol consumption | 62.5% |
| ADP29: Ever had problems with drug consumption | 72.8% |
| ADP30: Currently has problems with alcohol consumption | 39.0% |
| ADP31: Currently has problems with drug consumption | 48.8% |
| ADP32: Criminal behavior related to alcohol/drug consumption | 62.2% |
| ADP33: Problem with spouse/family related to alcohol/drug consumption | 41.0% |
| ADP34: Problem at school/work related to alcohol/drug consumption | 27.6% |
| ADP35: Health problems related to alcohol/drug consumption | 16.1% |
| Alcohol/Drug Problem total | 3.70 (2.36) |
| PA36: Favorable toward delinquency | 57.9% |
| PA37: Distrust toward society | 40.7% |
| PA38: Resentful toward sentence/offense | 37.6% |
| PA39: Uncooperative with supervision/treatment | 27.3% |
| Prcriminal Attitude/Orientation total | 1.64 (1.30) |
| AP40: High-risk mental health problem | 12.5% |
| AP41: Young and versatile delinquency | 38.0% |
| AP42: Antisocial values | 69.5% |
| AP43: Multiple problems | 54.3% |
| Antisocial pattern total | 1.74 (1.19) |
| LS/CMI total | 23.75 (8.62) |

Note. LS/CMI = Level of Service/Case Management Inventory; CH = Criminal History; EE = Education/Employment; FM = Family/Marital; LR = Leisure/Recreation; CO = Companions; ADP = Alcohol/Drug Problem; PA = Prcriminal Attitude/Orientation; AP = Antisocial Pattern.

*aLS/CMI items have been paraphrased to avoid copyright issues.*
Monotonicity. Monotonicity implies that an individual with a high ability on a latent trait should have a greater probability to endorse an item measuring that same trait than an individual with a lower ability. This postulate may be checked by plotting the observed data on each ICC to see if the probability to endorse an item increases with greater thetas. Monotonicity was checked post-Rasch modeling to ensure that data respected this assumption. The empirical curve was plotted against the theoretical curve and, although they did not stack perfectly, they were parallel and, thus, showed that an increase in ability yielded an increase in both the probability of endorsement and the observed proportion of endorsement of items. Monotonicity graphs can be found in the Supplementary Materials.

Unidimensionality. When using a unidimensional model, such as the Rasch model, evidence must be provided that a single or dominant trait is being measured (e.g., by studying the eigenvalues produced by a factor analysis of the data). A dominant trait can be assumed if a significant drop is seen between the first and second eigenvalues. Even though it was not the main purpose of our study, a factor analysis of the 43 LS/CMI items was conducted using MPlus 6.12 to assess if the scale was “unidimensional enough” for Rasch modeling (Bertrand & Blais, 2004). This analysis was based on the guidelines for risk tool factor analysis suggested by Brouillette-Alarie et al. (2016): (a) use of tetrachoric correlation matrices, (b) weighted least squares means- and variance-adjusted extraction, and (c) oblique (geomin) rotation. The first factor had 14.53 eigenvalues and the second 3.77, which constituted a ratio of 3.85 between the first and second factor. Usually, for a scale to be considered unidimensional enough for Rasch modeling, a ratio of 3 or higher is recommended (Bertrand & Blais, 2004). Thus, for the purposes of the Rasch model, the LS/CMI was considered sufficiently unidimensional.

Local independence. This postulate assumes that the success/failure on an item is independent of the success/failure on other items, thus solely dependent on the latent trait. Correlations are expected between items because, ideally, they all measure the same trait, but beyond that, there should not be excessively high correlations within the residuals. Because collinearity between LS/CMI items is notoriously high, we opted to not remove any items a priori and let the fit indices decide which ones would be ejected from the model. This also enabled us to obtain difficulty data on more items than if collinear items were discarded or summed beforehand.

Model-data fit. Conducting a Rasch analysis yields a difficulty parameter for each item and an ability parameter for each individual, with a standard error specific to each. The process also yields fit statistics for items and individuals, namely infit and outfit in mean square format. The outfit is calculated as the average of the squares of the standardized residuals (the residuals are squared before the averaging operation) and is particularly sensitive to the unexpected responses of people whose location is far from the item. Infit is calculated by multiplying the square of each standardized residual by the variance of the expected score and is particularly sensitive to responses expected from people whose location is close to the item. If data fit perfectly with model specifications, the expected values for infit and outfit are 1. While infit is rather hard to detect and interpret, the outfit is usually prominent (Linacre, 2002).
Different ranges of reasonable fit values are available depending on the nature of the test under scrutiny (Bond & Fox, 2001). Considering that the tool under study was the LS/CMI, we opted for the values that Wright and Linacre (1994) suggested for clinical observations. Thus, only items with fit statistics between 0.5 and 1.7 were kept. While Rasch calibration is usually done iteratively, each time removing the most unfitting item until all the remaining items fall within the desired range, the calibration only needed to loop once, since only one item was deemed unfitting under Wright and Linacre’s (1994) guidelines. The fit statistics for individuals were not scrutinized because response patterns were expected to vary.

For the sake of exhaustiveness, we also tried to run the Rasch model under strict conditions, namely those suggested for high-stakes tests (fit statistics between 0.8 and 1.2; Wright & Linacre, 1994). Under these circumstances, only 20 LS/CMI items remained after removing the unfitting ones. Because the aim of this study was to obtain data on LS/CMI items, discarding more than 50% of them a priori seemed counterproductive. Therefore, we settled on the more lenient clinical observation fit thresholds. All Rasch analyses were done with Winsteps (Linacre, 2021).

**Wright map.** The Wright map (also referred to as the item-person map) is a useful way to visualize both items and individuals vertically on the same graphic along the continuum of the targeted unidimensional space (Wright & Stone, 1979). A Wright map makes use of the fact that the difficulty of test items can be computed, and those test-item difficulties are expressed using the same linear scale as for the person measures. A logit scale is used to express item difficulty on a linear scale that extends from negative infinity to positive infinity. Item difficulty typically ranges from −3 (very easy) to +3 logits (very difficult; Boone et al., 2013). The Wright map depicts items organized according to difficulty level on the left and individuals positioned according to ability level on the right. In the Rasch model, the scale is set to zero for the item mean.

**Predictive Validity**

The predictive validity of LS/CMI items, dimensions, and total scores toward criminal recidivism was assessed with the AUC of receiver operating characteristic curves. AUCs refer to the probability that a randomly selected recidivist will have a higher score than a randomly selected nonrecidivist. It is an ordinal statistic that can be compared across different scaling of predictors. Rice and Harris’s (2005) thresholds for interpreting the effect sizes of AUCs were used: .556 is equivalent to a small effect, .639 is equivalent to a moderate effect, and .714 is equivalent to a large effect. These thresholds correspond, respectively, to Cohen’s $d_s$ of .20, .50, and .80. AUCs are statistically significant when their 95% confidence interval does not include .50.

**Interface Between Difficulty and Predictive Validity**

The link between the difficulty ($b$) and predictive validity (AUC) of items was obtained by computing the Pearson correlation between these two measures. The Pearson correlation was chosen because difficulty and predictive validity were normally distributed.
RESULTS

RASCH MODELING AND THE WRIGHT MAP

The Rasch model took one iteration beyond the initial one to have all items fit within the specified range (between 0.5 and 1.7, inclusively). Only Item 24 ended up being discarded. Table 2 shows the item parameters and fit statistics of the first and last iterations.

The 42 items were given a difficulty parameter, and the 15,961 incarcerated men were given an ability parameter through a joint maximum likelihood estimation. With all the item and person parameters estimated, the Wright map was drawn (see Figure 1), with the 15,961 participants on the left and the 42 items of the LS/CMI on the right.

The participants’ “ability” ($\Theta$) curve seemed to follow a normal distribution, slightly skewed negatively (to the left). Item 6 was the easiest item to endorse, while Item 40 was the hardest. The Wright map indicated which items targeted specific ranges of individuals. For instance, Items 10 and 43 were very well aligned with the ability of the “average” incarcerated individual ($\Theta \approx 0$), meaning that these items, when administered to these persons, generated the most desirable variance. Note that to avoid copyright issues, the names of LS/CMI items were paraphrased in the following sections.

The difficulty of individual items was ordered in the anticipated direction. For example, Item 3 (three prior convictions) was more difficult than Item 2 (two prior convictions), and the latter was more difficult than Item 1 (Any prior convictions). Item 30 (Currently has problems with alcohol consumption) was harder than Item 28 (Ever had problems with alcohol consumption). The Leisure/Recreation dimension was on average the easiest, followed by Criminal History. The two most difficult dimensions were Procriminal Attitude/Orientation and Antisocial Pattern. However, for the latter dimension, the mean difficulty was heavily influenced by Item 40, the most difficult item in our sample.

PREDICTIVE VALIDITY

The most predictive items were Items 7, 8, and 43, with moderate effect sizes. Items 6, 12, 13, 22, and 24 were not predictive in our sample. The remaining 34 items had small effect sizes. The most predictive dimension was Criminal History, which had a large effect size. The remaining seven dimensions had moderate effect sizes. No dimension had a negligible or small effect size. The total LS/CMI score had an AUC of .761, surpassing all the individual dimensions and items (its predictive validity was, however, close to that of the Criminal History dimension).

INTERFACE BETWEEN DIFFICULTY AND PREDICTIVE VALIDITY

The difficulty of each item (except CO24: Links with criminalized individuals) was correlated with its predictive validity to verify if difficult items were more predictive of recidivism than easier items. The scatter plot of LS/CMI items can be found in the Supplementary Materials, with predictive validity on the X-axis and difficulty on the Y-axis. The Pearson correlation between item difficulty and predictive validity was negligible ($r = .044, p = .781$), contradicting the assumption that more difficult items should be more predictive of the ability to commit crimes (recidivism). This was particularly illustrated in the Criminal History dimension, the most predictive but second to last dimension in terms of difficulty.
### Table 2: Rasch Modeling and Predictive Validity of LS/CMI Items (N = 15,961)

| LS/CMI items, dimensions, and total score<sup>a</sup> | First iteration | Second and last iteration | Predictive validity |
|------------------------------------------------------|-----------------|--------------------------|---------------------|
|                                                      | Diff. (b)       | Infit                     | Outfit              | AUC     |
| CH1: Any prior conviction                           | −2.14           | 0.89                      | 0.69                | .570*   |
| CH2: Two prior convictions                          | −1.38           | 0.88                      | 0.76                | .603*   |
| CH3: Three prior convictions                         | 0.86            | 0.91                      | 0.85                | .625*   |
| CH4: Three or more offenses for the current sentence | −0.79           | 1.15                      | 1.33                | .568*   |
| CH5: Arrested/charged before being 16 years old      | 1.57            | 0.97                      | 1.00                | .600*   |
| CH6: Ever incarcerated                               | −2.78           | 1.15                      | 1.28                | .531*   |
| CH7: Institutional misconduct                        | 0.62            | 0.94                      | 0.91                | .685*   |
| CH8: Breach of probation/parole conditions           | −1.01           | 0.85                      | 0.75                | .645*   |
| CH dimension Mean difficulty of items:               | −1.02 (SD = 1.40) |                           |                     | .742*   |
| EE9: Currently without a job                         | −0.46           | 0.91                      | 0.83                | .592*   |
| EE10: Frequently without a job                       | 0.22            | 0.87                      | 0.80                | .619*   |
| EE11: Never maintained a job for 1 year              | 1.80            | 0.95                      | 0.91                | .599*   |
| EE12: Did not achieve grade 10                       | 0.93            | 1.19                      | 1.33                | .546*   |
| EE13: Did not achieve grade 12                        | −0.83           | 1.18                      | 1.23                | .553*   |
| EE14: Suspended/expelled from school                 | 0.54            | 1.07                      | 1.12                | .604*   |
| EE15: Not motivated/successful at school/work        | −0.75           | 0.87                      | 0.75                | .592*   |
| EE16: Problematic relationships with peers at school/work | −0.59           | 0.89                      | 0.80                | .590*   |
| EE17: Problematic relationships with authority at school/work | −0.61           | 0.89                      | 0.79                | .590*   |
| EE dimension Mean difficulty of items:               | −0.14 (SD = 0.91) |                           |                     | .669*   |
| FM18: Not satisfied with intimate relationship or lack thereof | 1.31            | 1.18                      | 1.37                | .567*   |
| FM19: Problematic relationship with parents          | −0.04           | 1.20                      | 1.29                | .563*   |
| FM20: Problematic relationship with other relatives  | 0.71            | 1.07                      | 1.12                | .595*   |
| FM21: A family member or spouse has a criminal record | 0.77            | 1.22                      | 1.40                | .561*   |
| FM dimension Mean difficulty of items:               | 0.52 (SD = 0.56) |                           |                     | .639*   |
| LR22: No prosocial activities                        | −1.73           | 1.17                      | 1.32                | .545*   |
| LR23: Has too much free time                         | −0.50           | 0.89                      | 0.80                | .607*   |
| LR dimension Mean difficulty of items:               | −1.28 (SD = 0.86) |                           |                     | .618*   |
| CO24: Links with criminalized individuals            | −7.11           | 0.98                      | 0.41                | .501    |
| CO25: Friends with criminalized individuals          | 0.71            | 1.07                      | 1.11                | .579*   |
| CO26: Few prosocial links                           | 0.80            | 0.94                      | 0.91                | .600*   |
| CO27: Few prosocial friends                         | −0.88           | 0.97                      | 0.89                | .588*   |
| CO dimension Mean difficulty (without Item 24):       | 0.52 (SD = 0.36) |                           |                     | .645*   |
| ADP28: Ever had problems with alcohol consumption    | −0.23           | 1.21                      | 1.29                | .564*   |
| ADP29: Ever had problems with drug consumption       | −0.84           | 0.94                      | 0.84                | .605*   |
| ADP30: Currently has problems with alcohol consumption | 1.01            | 1.15                      | 1.17                | .575*   |
| ADP31: Currently has problems with drug consumption  | 0.50            | 0.90                      | 0.84                | .617*   |
| ADP32: Criminal behavior related to alcohol/drug consumption | −0.21           | 0.91                      | 0.84                | .614*   |
| ADP33: Problem with spouse/family related to alcohol/drug consumption | 0.91            | 0.99                      | 0.94                | .614*   |
| ADP34: Problem at school/work related to alcohol/drug consumption | 1.66            | 0.97                      | 0.89                | .579*   |
| ADP35: Health problems related to alcohol/drug consumption | 2.49            | 1.05                      | 1.04                | .556*   |
| ADP dimension Mean difficulty of items:              | 0.49 (SD = 1.09) |                           |                     | .677*   |
| PA36: Favorable toward delinquency                   | 0.02            | 0.96                      | 0.92                | .611*   |
| PA37: Distrust toward society                        | 0.92            | 0.94                      | 0.90                | .606*   |
| PA38: Resentful toward sentence/ offense             | 1.09            | 1.23                      | 1.38                | .556*   |
| PA39: Uncooperative with supervision/treatment       | 1.68            | 1.05                      | 1.04                | .576*   |
| PA dimension Mean difficulty of items:               | 0.76 (SD = 0.69) |                           |                     | .651*   |
| AP40: High-risk mental health problem                | 2.83            | 0.97                      | 0.92                | .566*   |
| AP41: Young and versatile delinquency                | 1.07            | 0.90                      | 0.85                | .631*   |
| AP42: Antisocial values                              | −0.63           | 0.88                      | 0.78                | .606*   |
| AP43: Multiple problems                              | 0.21            | 0.75                      | 0.66                | .644*   |
| AP dimension Mean difficulty of items:               | 0.70 (SD = 1.48) |                           |                     | .711*   |
| LS/CMI total score                                   |                 |                           |                     | .761*   |

<sup>a</sup>LS/CMI items have been paraphrased to avoid copyright issues.

Note. LS/CMI = Level of Service/Case Management Inventory; AUC = area under the curve; CH = Criminal History; EE = Education/Employment; FM = Family/Marital; LR = Leisure/Recreation; CO = Companions; ADP = Alcohol/Drug Problem; PA = Procriminal Attitude/Orientation; AP = Antisocial Pattern.

<sup>*</sup>AUCs are statistically significant when their 95% confidence interval does not include .50.
Figure 1: Wright Map of LS/CMI Items and Participants

Note. LS/CMI = Level of Service/Case Management Inventory; CH = Criminal History; EE = Education/Employment; FM = Family/Marital; LR = Leisure/Recreation; CO = Companions; ADP = Alcohol/Drug Problem; PA = Procriminal Attitude/Orientation; AP = Antisocial Pattern.
DISCUSSION

The objectives of the current study were to obtain item-level data on the LS/CMI and study the interface between IRT difficulty and predictive validity. The analyses conducted provided validation data on the French version of the LS/CMI with the whole population of Quebec’s incarcerated men registered and evaluated between March 2008 and October 2015.

LS/CMI ITEMS

Rasch model fit indices and predictive validity analyses highlighted potential problems in five items: Items 6, 12, 13, 22, and 24. For two of these items (6 and 24), sample characteristics were likely responsible for their poor performance. Indeed, according to the LS/CMI coding manual (Andrews et al., 2004), Item 6 (Ever incarcerated) and Item 24 (Links with criminalized individuals) must be endorsed for all individuals under custody. Because our sample exclusively comprised incarcerated persons, these items lacked variance, explaining their lack of predictive validity and, in the case of Item 24, exclusion from the Rasch model. If our sample had also comprised individuals under parole, the picture may have been different.

Items 12 (Did not achieve grade 10) and 13 (Did not achieve grade 12), both related to educational level, were not predictive of recidivism in our sample despite being adequately distributed. School dropout seemed like a better predictor than educational level, the same being true for Items 15, 16, and 17, which look, respectively, at performance, peer interactions, and authority interactions in schools (or jobs). It may be that problematic behaviors in school are better predictors of recidivism than educational level, which may confound multiple noncriminogenic characteristics, such as IQ, motivation, or learning style. A recent meta-analysis of risk factors for recidivism reached results similar to ours concerning the mediocre predictive validity of educational level (Goodley et al., 2022). As to Item 22 (No prosocial activities), there seemed to be no sample selection explanations for its lackluster predictive validity. We also found no studies specifically about the link between the absence of structured activities and general recidivism. Therefore, before making any conclusion concerning Item 22, the present findings would have to be replicated.

The two most predictive items were Items 7 and 8, which covered, respectively, institutional misconduct and breach of release conditions. The behaviors described by these items are known risk factors of general recidivism (Goodley et al., 2022) and figure in multiple criminological risk scales, such as the STABLE-2007 (Brankley et al., 2021; Hanson et al., 2007) and the PCL-R. In addition, an upcoming study based on machine learning algorithms concluded that these two items were the most predictive of general recidivism in a sample very similar to the one used in the current study (Arbour et al., 2022). However, this convergence of results may be partly explained by commonalities in the samples used.

The predictive validity of LS/CMI items was generally good in relation to field standards and what can be expected in predictive potency from single items. For comparison purposes, a meta-analysis of the predictive validity of Static-99R items toward sexual recidivism found that odds ratios varied between 1.22 and 2.47 (Helmus & Thornton, 2015). Odds ratios between 1.68 and 3.46 are considered small. Those between 3.47 and 6.70 are considered moderate, and those of 6.71+ are considered large (Chen et al., 2010). This would mean that for the Static-99R, the predictive validity of individual items varied
between negligible and small effects. For the LS/CMI in our sample, few items had negligible effects, most had small effects, and three items reached moderate effects. Considering that two of the nonpredictive items were because of sample selection, the overall picture of the predictive validity of individual LS/CMI items appeared adequate and in line with field standards, or better. We would, however, strongly encourage replication of these results, as more studies of the characteristics of individual LS/CMI items need to be conducted.

**LS/CMI DIMENSIONS AND TOTAL SCORE**

The most predictive dimension of the LS/CMI was Criminal History, a finding consistent with Olver et al.’s (2014) meta-analysis and the well-known reliability and predictive validity of static risk factors (e.g., Brouillette-Alarie & Lussier, 2018; Giguère & Lussier, 2016). The other dimensions all achieved moderate predictive validities that surpassed those of their constituents (items). These predictive validities were all superior to those reported by Olver et al. (2014) for the same dimensions.

The predictive validity of the LS/CMI total score for Quebec incarcerated men was in line with or better than effect sizes reported in other validation samples. In a literature review conducted by Olver et al. (2014), the predictive validity ($r$) of LS scales for general or violent recidivism ranged between .15 (Singh et al., 2011) and .39 (Gendreau et al., 2002). In our study, when converted into correlation metrics, the predictive validity of the LS/CMI for general recidivism was of .45. This level of predictive validity is rarely achieved by risk tools in the criminological field (see Campbell et al., 2009 or Langton et al., 2007 for comparisons). Thus, in relation to predictive validity, the LS/CMI performed admirably in our sample.

**RASCH DIFFICULTY AND ITS RELATIONSHIP WITH PREDICTIVE VALIDITY**

Rasch modeling provided the difficulty of the 42 LS/CMI items retained in the model. The two most difficult items ($b > 2.0$) were Item 40 (High-risk mental health problem) and Item 35 (Health problems related to alcohol/drug consumption). These items rely, among other things, on medical and/or psychiatric files (Andrews et al., 2004) which, according to professionals involved in the assessment of Quebec’s convicted individuals, were not always available. Because difficulty is heavily influenced by the percent of positive responses to an item, the difficulty of these items may have been overestimated due to the scarcity of files required to score this item. The easiest items ($b < -2.0$) were Item 6 (Ever incarcerated) and Item 1 (Any prior conviction), two items nearly automatically endorsed for incarcerated persons. Again, sample selection may have made these items easier than what would be anticipated in a sample comprising both incarcerated individuals and persons on parole.

The relative difficulty of items was in the anticipated direction, especially for items under the same dimension (e.g., Item 3 was more difficult than Item 2, which was more difficult than Item 1). It was, however, harder to contrast the difficulties of items not under the same dimension. For example, dissatisfaction with one’s marital situation (FM18) was more difficult than breaching one’s conditions during supervision (CH8). If the assumptions of the Rasch model are to be believed, the former should thus be more criminogenic than the latter. Predictive validity analyses revealed a different picture. Despite its significantly lower difficulty, Item 8 was far more predictive of general recidivism than Item 18. Rasch difficulty
and predictive validity were globally unrelated, which casts a doubt on some aspects of the usefulness of IRT techniques for risk tool validation. Despite the enthusiasm of some authors to completely discard CTT in favor of IRT to process criminogenic data (e.g., Osgood et al., 2002), the current study indicates that the difficulty parameter may not be useful to improve the predictive validity of actuarial scales. It may be that difficulty in mathematical tests from the field of education cannot be interpreted in the same way as difficulty in items from actuarial scales. For mathematical tests, success on a complicated item usually implies success on an easier item, as they rely on the same skill. However, in the context of risk scales, that assumption might not hold true. To echo the above example, it is unlikely that being dissatisfied with one’s marital situation (the harder item) “automatically comes” with breaching one’s parole conditions (the easier item), as these items relate to different dimensions or latent constructs.

The difficulty in interpreting LS/CMI item difficulties was exacerbated by the relative multidimensionality of the scale. Even though the LS/CMI was “unidimensional enough” to run Rasch models according to standards in the IRT field (Bertrand & Blais, 2004), it was not fully unidimensional in the factor analysis that was conducted. This finding aligns with more methodologically solid studies of the LS/CMI’s factor structure (e.g., Hsu et al., 2011), which also found numerous dimensions. The multidimensionality of recidivism risk was empirically demonstrated for persons convicted for a sexual offense (Brouillette-Alarie et al., 2016, 2018; Brouillette-Alarie & Hanson, 2015; Brouillette-Alarie & Proulx, 2019), and is implicit in the multiplicity of LS dimensions and the theoretical underpinnings of the psychology of criminal conduct (Andrews & Bonta, 2010). It may be that a higher-order “risk” construct encompasses the eight subscales of the LS/CMI, akin to the g factor in intelligence studies (e.g., Carroll, 1993; Johnson & Bouchard, 2005), but even then, as it relates to the analyses conducted here, the LS/CMI appeared to be measuring multiple latent traits. This limited the validity of comparing the difficulty scores of LS/CMI items. In contrast, IRT techniques have been applied to criminological scales measuring sexual sadism and have found results that are more easily interpretable due to clearer unidimensionality and a more limited number of items (Longpré et al., 2019; Mokros et al., 2012). It may be the case that for the LS/CMI, analyses that account for multidimensionality (e.g., multidimensional IRT) may prove more adapted to the task.

LIMITATIONS

Even though the current study aimed to be thorough in its methods and prudent in its conclusions, it is not without limitations. First, the results reported here are limited to incarcerated men from Quebec’s provincial prison system, meaning that our sample, even though it was a population, is not necessarily representative of other populations. Specifically, it may not be generalizable to (a) potentially higher-risk persons from Canadian federal prisons that have received a sentence of 2 years or more; (b) individuals who have received a community sentence or are on parole; (c) women; (d) ethnoculturally diverse correctional populations; and (e) individuals involved in the US correctional system—especially in light of the lackluster predictive validity found in evaluation studies of the LS/CMI in the US.

Second, as mentioned above, the potential multidimensionality of the LS/CMI, as well as its high number of items, may have curtailed the usefulness of Rasch modeling in this study. However, rather than sweeping the issue under the rug, we thoroughly discussed it by
plotting the difficulty of items by their predictive validity. As such, results and discussions from the current study may be of use to future scholars who try to apply IRT techniques to criminological risk scales. Third, the local independence assumption could not be fully met as removing or merging collinear items from the LS/CMI would have resulted in merging nearly half of the items, which would have deprived readers of valuable data on LS/CMI items. In future studies, especially for analyses highly sensitive to collinearity (e.g., factor analysis), more thorough item preparation may be necessary. Fourth, the follow-up period of the current study was limited to 2 years, which may leave “little time” for individuals to reoffend, especially for individuals who are on the lower end of the risk spectrum. Finally, although we do not anticipate that such a limitation would have significant effects on our results, it is worthwhile to mention that the conclusions of the current study are based on the French version of the LS/CMI and may thus not be applicable to the English version of the scale.

**IMPLICATIONS FOR RESEARCH AND PRACTICE**

Taken together, results of the current study attest to the use of the LS/CMI to assess the risk of general recidivism in incarcerated men from the Quebec population. The predictive validity of LS/CMI items, dimensions, and total scores was very good and among the best of what the field has to offer. Even under Rasch scrutiny, items performed relatively well and covered the whole spectrum of the risk continuum (difficulties ranging from −2.94 to 2.66). Apart from items that suffered from sample selection, few items had lackluster effect sizes toward general recidivism.

The current study offers some interesting avenues for future developments of the LS/CMI in relation to data that was obtained concerning its items. First, the substantial variation in the predictive validity and difficulty of items challenges the equal weight (one point) attributed to all LS/CMI items. However, contrary to our initial hypotheses, item difficulty was thoroughly unrelated to predictive validity. As such, results from the current study indicate that the predictive validity of items may be a better basis upon which to re-weight items than Rasch difficulty. Second, should the lack of predictive validity of items related to educational level be replicated in other samples, it might be worthwhile to rethink their presence in the scale, especially in light of meta-analyses that challenged the association between educational level and delinquency (Goodley et al., 2022). Third, predictive validity analyses revealed that some items (7 and 8) were particularly indicative of recidivism potential. Because these items refer, respectively, to institutional misconduct and parole breach, correctional staff should be particularly wary of these behaviors as they may indicate relapse. Even though the LS/CMI does not explicitly integrate the stable/acute distinction, we anticipate that Items 7 and 8 would be prime candidates for acute risk factors of general recidivism. Fourth, the factor analysis that was made to ensure “sufficient unidimensionality” challenged the unidimensionality of the LS/CMI put forward by many authors (see Hsu et al., 2011). We think that robust factor analytic studies of such a widely used scale are overdue and could enable a comparison between the eight theoretical dimensions of the scale and its empirical latent structure.

Finally, as it relates to fundamental research, the current study highlighted some limitations of applying Rasch models from the field of education to risk assessment scales from the field of criminology. Rasch modeling has shown that items deemed difficult to endorse
were not necessarily risk relevant and did not systematically show good predictive validity. Even though this could dampen the enthusiasm of researchers toward IRT-based models, IRT does offer some underexplored options for the psychometric validation of risk scales. Namely, IRT techniques could enable the investigation of the usefulness of the discrimination parameter (instead of difficulty) to improve predictive validity. In addition, using differential item functioning could help to determine if items perform equally well for different groups of convicted persons (e.g., men vs. women). Importantly, this technique could elucidate other important psychometric properties of risk scales for Indigenous populations, which have been the focus of increased legal challenge and scrutiny in recent years (Gutierrez et al., 2017; Huang et al., 2021).

ORCID ID

Sébastien Brouillette-Alarie https://orcid.org/0000-0002-7620-8786

SUPPLEMENTAL MATERIAL

Supplemental Material is available in the online version of this article at http://journals.sagepub.com/home/cjb

REFERENCES

Andrews, D. A., & Bonta, J. (1995). *The Level of Service Inventory—Revised*. Multi-Health Systems.
Andrews, D. A., & Bonta, J. (1998). *The Level of Service Inventory—Revised: Screening version*. Multi-Health Systems.
Andrews, D. A., & Bonta, J. (2010). *The psychology of criminal conduct* (5th ed.). LexisNexis/Matthew Bender.
Andrews, D. A., Bonta, J., & Wormith, J. S. (2004). *Manual for the Level of Service/Case Management Inventory (LS/CMI)*. Multi-Health Systems.
Andrews, D. A., Bonta, J., Wormith, J. S., Guzzo, L., Brews, A., Rettinger, J., & Rowe, R. (2011). Sources of variability in estimates of predictive validity: A specification with level of service general risk and need. *Criminal Justice and Behavior*, 38(5), 413–432. https://doi.org/10.1177/0093854811401990
Andrews, D. A., & Robinson, D. (1984). *The Level of Supervision Inventory: Second report*. Ontario Ministry of Correctional Services.
Arbour, W., Brouillette-Alarie, S., Giguère, G., Lacroix, G., & Marchand, S. (2022). *Predicting (and misclassifying) offenders’ risk*. Manuscript submitted for preparation.
Arens, S. A., Durham, B., O’Keefe, M., Klebe, K., & Olene, S. (1996). *Psychometric properties of Colorado substance abuse assessment instruments*. Unpublished manuscript.
Babchishin, K. M., & Hanson, R. K. (2020). Monitoring changes in risk of reoffending: A prospective study of 632 men on community supervision. *Journal of Consulting and Clinical Psychology*, 88(10), 886–898. https://doi.org/10.1037/ccc0000601
Babchishin, K. M., Hanson, R. K., & Blais, J. (2016). Less is more: Using Static-2002R subscales to predict violent and general recidivism among sexual offenders. *Sexual Abuse: A Journal of Research and Treatment*, 28(3), 187–217. https://doi.org/10.1177/1079063215569544
Barbaree, H. B., Langton, C. M., Blanchard, R., & Cantor, J. (2009). Aging versus stable enduring traits as explanatory constructs in sex offender recidivism: Partitioning actuarial prediction into conceptually meaningful components. *Criminal Justice and Behavior*, 36(5), 443–465. https://doi.org/10.1177/0093854809332283
Bertrand, R., & Blais, J.-G. (2004). *Modèles de mesure: L’apport de la théorie des réponses aux items* [Measurement models: The contribution of item response theory]. Presses de l’Université du Québec.
Bohlig, M., Fisher, W. P., Jr., Masters, G. N., & Bond, T. (1998). Content validity and unfitting items. *Rasch Measurement Transactions*, 12(1), 607. https://www.rasch.org/rmt/rmt121f.htm
Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Lawrence Erlbaum.
Bonta, J. (1996). Risk-needs assessment and treatment. In A. T. Harland (Ed.), *Choosing correctional options that work: Defining the demand and evaluating the supply* (pp. 18–32). SAGE.
Bonta, J. (2002). Offender risk assessment: Guidelines for selection and use. *Criminal Justice and Behavior*, 29(4), 355–379. https://doi.org/10.1177/0093854802029004002
Boone, W. J. (2016). Rasch analysis for instrument development: Why, when, and how? *CBE Life Sciences Education*, 15(4), rm4. https://doi.org/10.1187/cbe.16-04-0148
Boone, W. J., Staver, J. R., & Yale, M. S. (2013). *Rasch analysis in the human sciences*. Springer.
Brouillette-Alarie, S., Babchishin, K. M., & Hanson, R. K. (2016). STABLE-2007 demonstrates predictive and incremental validity in assessing risk-relevant propensities for sexual offending: A meta-analysis. *Sexual Abuse, 33*(1), 34–62. doi.org/10.1177/1079063218759325

Brouillette-Alarie, S., Bouchard, M. F., Bub, A. E., Babchishin, K. M., & Hanson, R. K. (2019). Three central dimensions of sexual recidivism risk: Understanding the latent constructs of Static-99R and Static-2002R. *Sexual Abuse, 30*(6), 676–704. doi.org/10.1177/1079063217769165

Campbell, M. A., French, S., & Gendreau, P. (2009). The prediction of violence in adult offenders: A meta-analytic comparison of instruments and methods of assessment. *Criminal Justice and Behavior, 36*(6), 567–590. doi.org/10.1177/0093854809333610

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press.

Chen, H., Cohen, P., & Chen, S. (2010). How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies. *Communications in Statistics—Simulation and Computation, 39*(4), 860–864. doi.org/10.1080/03610911003650383

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. doi.org/10.1037/h0040957

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*(4899), 1668–1674. https://doi.org/10.1126/science.2648573

de Ayala, R. J. (2009). *Methodology in the social sciences. The theory and practice of item response theory*. Guilford Press.

Engelhard, G., & Wang, J. (2021). Rasch models for solving measurement problems: Invariant measurement in the social sciences. *SAGE.

Gendreau, P., Goggin, C., & Smith, P. (2002). Is the PCL-R really the “unparalleled” measure of offender risk? A lesson in knowledge cumulation. *Criminal Justice and Behavior, 29*(4), 397–426. doi.org/10.1177/0093854802029004004

Gendreau, P., Little, T., & Goggin, C. (1996). A meta-analysis of the predictors of adult offender recidivism: What works? *Criminology, 34*(4), 575–608. doi.org/10.1111/j.1745-9125.1996.tb01220.x

Georgiou, G. (2019). Weights matter: Improving the predictive validity of risk assessments for criminal offenders. *Journal of Offender Rehabilitation, 58*(2), 92–116. doi.org/10.1080/10509674.2018.1562504

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press.

Giguère, G., & Lussier, P. (2016). Debunking the psychometric properties of the LS/CMI: An application of item response theory with a risk assessment instrument. *Journal of Criminal Justice, 46*, 207–218. doi.org/10.1016/j.jcrimjus.2016.05.005

Giguère, G., Savard, D., & Cortoni, F. (2015). Une étude psychométrique des items du Level of Service / Case Management Inventory (LS/CMI) avec la contribution de la théorie classique des tests chez les personnes contrevenantes du Québec [A psychometric study of LS/CMI items using classical test theory in Quebec offenders]. *Canadian Journal of Criminology and Criminal Justice, 57*(3), 293–329. doi.org/10.3138/cjccj.2013.F06

Goodley, G., Pearson, D., & Morris, P. (2022). Predictors of recidivism following release from custody: A meta-analysis. *Psychology, Crime & Law, 28*(7), 703–729. doi.org/10.1080/10790632.2021.1962866

Gordon, H., Kelty, S. F., & Julian, R. (2015). Psychometric evaluation of the Level of Service/Case Management Inventory among Australian offenders completing community-based sentences. *Criminal Justice and Behavior, 42*(11), 1089–1109. doi.org/10.1177/0093854815596419

Grann, M., & Långström, N. (2007). Actuarial assessment of violence risk: To weigh or not to weigh? *Criminal Justice and Behavior, 34*(1), 22–36. doi.org/10.1177/0093854806290250

Grove, W. M., Zald, D. H., Lewob, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12*(1), 19–30. doi.org/10.1037/1040-3590.12.1.19

Guay, J.-P. (2016). L’évaluation du risque et des besoins criminogènes à la lumière des données probantes: Une étude de validation de la version Française de l’inventaire de niveau de service et de gestion des cas—LS/CMI [French validation of the Level of Service/Case Management Inventory—LS/CMI]. *European Review of Applied Psychology/Revue Européenne de Psychologie Appliquée*, 66(4), 199–210. doi.org/10.1016/j.erap.2016.04.003
Osteen, P. (2010). An introduction to using multidimensional item response theory to assess latent factor structures. *Journal of the Society for Social Work and Research, 1*(2), 66–82. https://doi.org/10.5243/jsswr.2010.6

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.

Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC Area, Cohen's d, and r. *Law and Human Behavior, 29*(5), 615–620. https://doi.org/10.1007/s10979-005-6832-7

Silver, E., Smith, W. R., & Banks, S. (2000). Constructing actuarial devices for predicting recidivism: A comparison of methods. *Criminal Justice and Behavior, 27*(6), 733–764. https://doi.org/10.1177/0093854800027006004

Singh, J. P., Gramm, M., & Fazel, S. (2011). A comparative study of violence risk assessment tools: A systematic review and metaregression analysis of 68 studies involving 25,980 participants. *Clinical Psychology Review, 31*(3), 499–513. https://doi.org/10.1016/j.cpr.2010.11.009

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. https://doi.org/10.1080/14999013.2011.577138

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370. www.rasch.org/rmt/rmt83b.htm

Wright, B. D., & Stone, M. (1979). *Best test design*. MESA Press.

**Guy Giguère** is a Senior Researcher for the Ministère de la Sécurité publique du Québec (Canada) and a lecturer at the Université Laval and the Université du Québec à Trois-Rivières. He holds a Ph.D. in Measurement and Evaluation from the Université Laval and completed a postdoctoral fellowship at the International Center for Comparative Criminology (Université de Montréal). He specializes in psychometry and risk assessment.

**Sébastien Brouillette-Alarie**, Ph.D., is an Associate Professor and Lecturer at the École de Criminologie of the Université de Montréal. He is also the Scientific Coordinator of the Canadian Practitioners Network for the Prevention of Radicalization and Extremist Violence (CPN-PREV), located at the Université du Québec à Montréal. His areas of expertise are sexual violence, violent radicalization, risk assessment, and psychometry.

**Christian Bourassa** is a research professional at the Université de Montréal, where he is pursuing a Ph.D. in education measurement and statistics under the supervision of Sébastien Béland and Christophe Chénier. He specializes in adaptive testing and item response theory.