p-ISSN: 2184-2205 e-ISSN: 2184-3414 ISBN: 978-989-36839-1-0 © 2026
https://doi.org/10.36315/2026inpact046

RISK ASSESSMENT IN PSYCHOLOGY: PAST MISTAKES AND FUTURE DIRECTIONS FOR RESEARCH

Johannes Woschizka, Martina Feldhammer-Kahr, & Martin Arendasy
Department of Psychology, University of Graz (Austria)

Abstract

Forensic risk assessment has been a cornerstone of psychological research and practice since the beginning of the 20th century. Early risk appraisals relied on clinical judgment to predict outcomes like violent or criminal behavior, but these methods often lacked accuracy. The subsequent shift to actuarial approaches introduced empirically derived tools that aimed to provide objective assessments, thereby minimizing the subjectivity of clinical judgment and increasing accuracy. However, these actuarial methods were criticized for being atheoretical and failing to address criminogenic risks and needs, which became increasingly relevant as penal systems emphasized the resocialization of incarcerated individuals. In response, risk and need assessments and structured professional judgment (SPJ) tools emerged, integrating empirically established correlates of criminal behavior with clinical expertise. Despite these advancements, persistent methodological and theoretical issues remain largely unaddressed. This review critically reexamines the evolution of forensic risk assessment, highlighting a key shortcoming in existing approaches to risk assessment. We propose a roadmap for future research to enhance the validity, applicability and ethical use of risk assessment instruments in forensic psychology.

Keywords: *Forensic risk assessment, latent variable modeling, psychometrics, recidivism.*

1. Introduction

Risk assessment is a core component of the daily work forensic psychologists and psychiatrists perform. Especially the last twenty-five years are best characterized by an increasing demand for evidence-based risk-assessment instruments from the behavioral and social sciences (Monahan, 2018). However, concerns regarding the development, validation (Feldhammer-Kahr, Kaiser, Leibetseder, & Arendasy, 2024; Taxman, 2018) and a lack of reporting standards (e.g., Singh, Yang, Mulvey, & RAGEE Group, 2015), as well as the implementation into numerous areas of the justice system have sparked controversial discussions about the adequacy of the application of such instruments (Berk, Heidari, Jabbari, Kearns, & Roth, 2017; Hannah-Moffat, 2013, 2019). Despite the breadth of critical perspectives, what remains is a noticeable underrepresentation of perspectives informed by psychometric theory.

This article will first introduce forensic risk assessment in psychology before presenting a historical perspective on the development of risk and need assessment instruments, highlighting the significant improvements and changes throughout this development. The next part will then show that persistent issues are rooted in inconsistent application and evaluation of psychometric models, which disrupt theoretical advancements. Lastly, future directions towards a more psychometrically sound evaluation of theoretically derived risk and need assessment instruments are discussed.

2. The History of forensic risk assessment instruments

As commonly defined, forensic risk assessment in psychology is a diagnostic process in which information is gathered, weighed and synthesized, supported by standardized psychological risk assessment instruments (Bonta, 2002). Besides the voiced concerns of purely actuarial risk assessment (Harcourt, 2007), there exists a solid body of evidence that highlights the benefits of instrument-assisted risk assessment for predictive accuracy (e.g., Ægisdóttir, et al., 2006). These benefits should, however, not distract us from persistent issues of current risk assessment instruments.

In the past, four approaches of risk assessment – so called generations – have been identified (Andrews, Bonta, & Wormith, 2006). In the first generation of risk assessments, forensic professionals were tasked with estimating the risk of recidivism for individuals based on their professional expertise and commonly assisted by (unstructured) interviews.

Due to criticism regarding the explanatory power, inconsistency and a lack of transparency of unstructured assessments, a more empirically guided approach was developed (Burrell, 2017). Actuarial instruments assessed empirically derived static risk factors such as prior criminal offenses. They subsequently were criticized because offenders were not able to reduce risk scores through compliance with treatments or rehabilitation programs. Examples of widely used instruments include the Psychopathy Checklist-Revised (PCL-R; Hare, 1991) and the Violence Risk Appraisal Guide-Revised (VRAG-R; Rice, Harris & Lang, 2013).

To address these limitations, dynamic risk factors were included in the third generation of instruments (Burrell, 2017). This generation of instruments is referred to as "risk/needs scales" (Andrews & Bonta, 2014). Criminogenic needs are dynamic risk factors that can be changed through intervention and result in a lower likelihood of adverse behaviors. The third generation also includes structured professional judgements (SPJ), which are instruments that 'adopt non-algorithmic, non-numeric decision processes and risk estimates' (Douglas et al., 2014, p. 94). SPJ offer a structured way of assessing risk while enabling clinical experience and expertise to influence the decision (Fletcher, Gredecki, & Turner, 2021). Third generation instruments (especially SPJs) are among the most popular forensic risk assessment instruments (Singh et al., 2014). The HCR-20 and the Level of Service Inventory-Revised (LSI-R; Andrews & Bonta, 1995) are the most widely used instruments in practice.

The latest generation of risk assessment instruments enables practitioners to include case management, and individual differences in motivation and responsivity in their assessment (Andrews & Bonta, 2014; Fletcher et al., 2021). The Level of Service/Case Management Inventory (LS/CMI; Andrews, Bonta, & Wormith, 2004) is a fourth-generation instrument that includes additional protective factors, which are not considered for the calculation of risk scores. Another example is the empirically derived Static Risk Offender Need Guide for Recidivism (STRONG-R; Hamilton et al., 2016).

3. The evolution of risk assessments from a psychometric perspective

The evolution from unstructured clinical judgement to static actuarial risk assessment that subsequently led to the inclusion of criminogenic needs and the development of SPJs, case management systems, and responsivity assessments is commonly framed as a success story (e.g., Andrews et al., 2006). Although this success narrative generally seems justified, psychometric evaluations of risk assessment instruments reveal a more nuanced conclusion. As Taxman (2018, p. 282) notes: "risk assessment tools are developed and implemented in the field without attention to many of the reliability and predictive validity issues that have affected the current state of the art". We therefore propose that the history of risk assessment instruments should be reconsidered from a psychometric point of view. We will discuss predictive and construct validity because they reveal two central issues of risk assessment instruments.

3.1. Predictive validity

Predictive validity is a numerical estimate of how well risk scores or categories predict actual adverse behavior. Predictive validity is assessed with area under the curve (AUC) statistics (Rice & Harris, 2005). AUCs summarize the performance of classification tasks by relating the true positive rate to the false positive rate at every possible threshold. AUC scores in the context of RNAs represent the probability that a randomly drawn score from a recidivist population is higher than a randomly drawn score from a non-recidivist population. However, AUCs only assess how well an instrument can differentiate between recidivists and non-recidivists (Singh, 2013), which does not quantify the relation between the predicted level of risk and the observed risk. Systematic over- or underestimations of risk can result in high discriminative validity estimates (AUCs) but low calibration.

As prior reviews have shown, only marginal improvements have been made since the second generation in terms of predictive validity (Fazel, Singh, Doll, & Grann, 2012; Singh, Grann, & Fazel, 2011; Yang, Wong, & Coid, 2010). We, among others, argue is that a strict focus on predictive validity has some major downsides for a thorough evaluation of RNAs. Usually, total scores or subscale scores are used to predict adverse outcomes. This process in itself assumes that subscales or test scores represent a single construct (such as "criminal history" or "risk of reoffending") that would justify the calculation of total scale scores (assumption of dimensionality). Further, if we presume that this assumption of dimensionality was tested and satisfied, the predictive validity of total scores does not provide information about the structural relations among predictors. The probability of reoffending in two populations of offenders may be affected differently by specific subscales, even if the predictive validity of the total score may be similar

in both populations. Differences in the relation of subscale scores to probabilities of reoffending contain relevant information for the prevention of future crimes.

3.2. Construct validity

Construct validity describes how measured indicators relate to a latent unmeasured concept of risk. Risk has been conceptualized as either an effect or causal indicator within risk assessment instruments. Effect indicators (or reflective factors) represent a common underlying unmeasured latent construct or cause of certain behaviors (Strube, 2015). Causal indicators (or formative factors) follow the conceptual idea that risk is best defined by indicators (such as criminal history) that represent facets of the latent construct. For example, effect indicators (e.g., intelligence) are the cause of answers given on a test, while causal indicators (e.g., socioeconomic status) reverse this relationship with answers or indicators defining the construct (e.g., household income, among other indicators). This difference has critical consequences for the evaluation of both types of testing procedures. At this point, it is important to note that although many risk assessment instruments have not specified a measurement model nor have been sufficiently evaluated with the appropriate methods, they presuppose a measurement model as soon as numerals are assigned to items and summarized into test scores.

The Violence Risk Appraisal Guide–Revised (VRAG-R; Rice et al., 2013), among other second-generation instruments such as the Risk Matrix 2000 (RM2000; see Thornton et al., 2003), conceptualizes risk as a causal indicator (Helmus & Quinsey, 2020).

Third-generation instruments on the other hand like the LSI-R and the Sexual Violence Risk-20 (SVR-20v2: Boer, Hart, Kropp, & Webster, 2017) tried to improve the second-generation instruments by adding criminogenic needs to risk assessment instruments. Although criminogenic needs are different from static risk factors, both get summarized as a total risk score. However, when researchers add a new dimension to the estimation of risk, they ought to show that empirical data allow for the recreation of theoretically proposed structures. Specifically, it must be shown that criminogenic needs and static risk are related (in some form, dependent on the proposed model) to the same latent risk. Investigations into the factor structure of widely used third-generation risk assessment instruments are sparse and have yielded inconsistent results that contradict theoretical models (Hsu, Caputi, & Byrne, 2011; Zhang & Liu, 2015). Although these methods are not directly applicable to SPJs, psychometric evaluations have revealed similar inconsistencies (Kanters et al., 2017; Klepfisz, Daffern, Day, Lloyd, & Woldgabreal, 2020).

A more nuanced picture arises for fourth-generation RNAs such as the LS/CMI and the STRONG-R (Hamilton et al., 2016). Similar to the LSI-R, inconsistent results have been reported for the LS/CMI (Giguère & Lussier, 2016). Contrary to other instruments, investigations into the structure of the STRONG-R have revealed an adequate factor structure of five subscales and a general risk factor (Hamilton et al., 2025). By explicitly excluding static risk factors, the authors provide ample evidence to justify the calculation of summary risk scores, including structural stability across gender and ethnic groups and negligible predictive bias (Hamilton et al., 2025). Although the STRONG-R provides an exemplary template for future development and validation practices, it remains to be seen if the results are replicable outside of the validation sample.

4. Conclusion

As the overview of the literature indicates, substantial inconsistencies exist regarding psychometric measurement models used by developers of forensic risk and need assessment instruments. While second-generation instruments rely on simple causal indicators, which were criticized for being atheoretical, third- and fourth-generation instruments argue that risk is better conceptualized as a multidimensional effect indicator. Most third and fourth-generation instruments have not provided sufficient psychometric evidence to support this claim. To illustrate this more clearly, Kroner, Mills, and Reddon (2005) have shown that similar predictive validity could be achieved by randomly restructuring the items of four popular risk assessment instruments. The authors conclude that according to their results, no single instrument has captured sufficient aspects of a theoretically superior risk assessment theory and all assessments capture a general construct of criminal risk (Kroner et al., 2005). This issue is not merely academic. When total scores are computed without a validated measurement model and prior assessment of dimensionality, they risk capturing a broad and unstable concept of risk. If interventions, parole decisions and rehabilitative efforts are based on the unstable concept of risk, they can potentially misallocate scarce resources and negatively affect the effectiveness of justice systems.

The issues raised in this paper require further investigation and substantiation with empirical data. What can be said at this point, however, is that further development of risk assessment instruments will require reconceptualization and modernization of validation procedures. This includes evaluations using latent variable models and comprehensive assessment of predictive validity. Researchers and test

developers should further be transparent about their risk concept about whether their risk concept is modeled with causal or effect indicators and evaluate it accordingly. Construct validity should be treated as a precondition of predictive validity, which especially applies to effect indicators. Risk assessment instruments can then also be more meaningfully evaluated in terms of their predictive validity, including a substantiation of discrimination with calibration indices. Rethinking validation in this way can align psychometric principles with the specific requirements of forensic risk assessment, resulting in more reliable decision-making processes and ethical implementation of instruments in practice.

References

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

Andrews, D. A., & Bonta, J. (2014). *The Psychology of Criminal Conduct* (5th ed). Taylor and Francis.

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. (2006). The Recent Past and Near Future of Risk and/or Need Assessment. *Crime & Delinquency*, *52*(1), 7–27. https://doi.org/10.1177/0011128705281756

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., Nichols, C. N., Lampropoulos, G. K., Walker, B. S., Cohen, G., & Rush, J. D. (2006). The Meta-Analysis of Clinical Judgment Project: Fifty-Six Years of Accumulated Research on Clinical Versus Statistical Prediction. *The Counseling Psychologist*, *34*(3), 341–382. https://doi.org/10.1177/0011000005285875

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2017). Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research, 50*. https://doi.org/10.1177/0049124118782533

Boer, D. P., Hart, S. D., Kropp, P. R., & Webster, C. D. (2017). *Manual for version 2 of the sexual violence risk–20: Structured professional judgment guidelines for assessing and managing risk of sexual violence.* Protect International Risk and Safety Services Inc.

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

Burrell, W. D. (2017). Risk and Needs Assessment in Probation and Parole: The Persistent Gap Between Promise and Practice. In F. S. Taxman (Ed.), *Handbook on risk and need assessment: Theory and practice*. Routledge.

Douglas, K. S., Hart, S. D., Webster, C. D., Belfrage, H., Guy, L. S., & Wilson, C. M. (2014). Historical-Clinical-Risk Management-20, Version 3 (HCR-20V3): Development and Overview. *International Journal of Forensic Mental Health*, *13*(2), 93–108. https://doi.org/10.1080/14999013.2014.906519

Fazel, S., Singh, J. P., Doll, H., & Grann, M. (2012). Use of risk assessment instruments to predict violence and antisocial behaviour in 73 samples involving 24 827 people: Systematic review and meta-analysis. *BMJ, 345*, e4692. https://doi.org/10.1136/bmj.e4692

Feldhammer-Kahr, M., Kaiser, N., Leibetseder, I., & Arendasy, M. (2024). Risk Appraisal and Legal Principles: Unveiling disciplinary gaps. In *Advances in Psychology and Psychological Trends Series: X*. inScience Press. https://doi.org/10.36315/2024padX25

Fletcher, M., Gredecki, N., & Turner, P. (2021). Forensic risk assessments. In *Forensic Psychology*. Routledge.

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. https://doi.org/10.1016/j.jcrimjus.2016.05.005

Hamilton, Z., Kigerl, A., Campagna, M., Barnoski, R., Lee, S., Van Wormer, J., & Block, L. (2016). Designed to Fit: The Development and Validation of the STRONG-R Recidivism Risk Assessment. *Criminal Justice and Behavior*, *43*(2), 230–263. https://doi.org/10.1177/0093854815615633

Hamilton, Z., Mei, X., Tostlebe, J. J., Allen-Flores, B., Ursino, J., & Kigerl, A. (2025). A methodological template for the next generation: Redesigning the STRONG-R needs assessment. *Journal of Criminal Justice*, *99*, 102454. https://doi.org/10.1016/j.jcrimjus.2025.102454

Hannah-Moffat, K. (2013). Actuarial Sentencing: An "Unsettled" Proposition. *Justice Quarterly*, *30*(2), 270–296. https://doi.org/10.1080/07418825.2012.682603

Hannah-Moffat, K. (2019). Algorithmic risk governance: Big data analytics, race and information activism in criminal justice debates. *Theoretical Criminology*, *23*(4), 453–470. https://doi.org/10.1177/1362480618763582

Harcourt, B. E. (2007). *Against prediction: Profiling, policing, and punishing in an actuarial age*. Chicago: University of Chicago Press.

Hare, R. D. (1991). *The Hare Psychopathy Checklist—Revised*. Multi-Health Systems.

Helmus, L. M., & Quinsey, V. L. (2020). Predicting Violent Reoffending with the VRAG-R: Overview, Controversies, and Future Directions for Actuarial Risk Scales. In J. S. Wormith, L. A. Craig, & T. E. Hogue (Eds.), *The Wiley Handbook of What Works in Violence Risk Management* (1st ed., pp. 119–144). Wiley. https://doi.org/10.1002/9781119315933.ch6

Hsu, C.-I., Caputi, P., & Byrne, M. K. (2011). The Level of Service Inventory-Revised (Lsi-R) and Australian Offenders: Factor Structure, Sensitivity, and Specificity. *Criminal Justice and Behavior*, *38*(6), 600–618. https://doi.org/10.1177/0093854811402583

Kanters, T., Hornsveld, R. H. J., Nunes, K. L., Zwets, A. J., Muris, P., & Van Marle, H. J. C. (2017). The Sexual Violence Risk-20: Factor structure and psychometric properties. *The Journal of Forensic Psychiatry & Psychology*, *28*(3), 368–387. https://doi.org/10.1080/14789949.2017.1284887

Klepfisz, G., Daffern, M., Day, A., Lloyd, C. D., & Woldgabreal, Y. (2020). Latent constructs in the measurement of risk and protective factors for violent reoffending using the HCR-20v3 and SAPROF: Implications for conceptualizing offender assessment and treatment planning. *Psychology, Crime & Law*, *26*(1), 93–108. https://doi.org/10.1080/1068316X.2019.1634197

Kroner, D. G., Mills, J. F., & Reddon, J. R. (2005). A Coffee Can, factor analysis, and prediction of antisocial behavior: The structure of criminal risk. *International Journal of Law and Psychiatry*, *28*(4), 360–374. https://doi.org/10.1016/j.ijlp.2004.01.011

Monahan, J. (2018). Recidivism Risk Assessment in the 21st Century. In J. P. Singh, D. G. Kroner, J. S. Wormith, S. L. Desmarais, & Z. Hamilton (Eds.), *Handbook of Recidivism Risk / Needs Assessment Tools*. John Wiley & Sons, Incorporated.

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

Rice, M. E., Harris, G. T., & Lang, C. (2013). Validation of and revision to the VRAG and SORAG: The Violence Risk Appraisal Guide—Revised (VRAG-R). *Psychological Assessment*, *25*(3), 951–965. https://doi.org/10.1037/a0032878

Singh, J. P. (2013). Predictive Validity Performance Indicators in Violence Risk Assessment: A Methodological Primer. *Behavioral Sciences & the Law*, *31*(1), 8–22. https://doi.org/10.1002/bsl.2052

Singh, J. P., Desmarais, S. L., Hurducas, C., Arbach-Lucioni, K., Condemarin, C., Dean, K., Doyle, M., Folino, J. O., Godoy-Cervera, V., Grann, M., Ho, R. M. Y., Large, M. M., Nielsen, L. H., Pham, T. H., Rebocho, M. F., Reeves, K. A., Rettenberger, M., De Ruiter, C., Seewald, K., & Otto, R. K. (2014). International Perspectives on the Practical Application of Violence Risk Assessment: A Global Survey of 44 Countries. *International Journal of Forensic Mental Health*, *13*(3), 193–206. https://doi.org/10.1080/14999013.2014.922141

Singh, J. P., Grann, 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

Singh, J. P., Yang, S., Mulvey, E. P., & The RAGEE Group. (2015). Reporting guidance for violence risk assessment predictive validity studies: The RAGEE Statement. *Law and Human Behavior*, *39*(1), 15–22. https://doi.org/10.1037/lhb0000090

Strube, M. J. (2015). Effect Indicator Versus Causal Indicator Measurement. In R. L. Cautin & S. O. Lilienfeld (Eds.), *The Encyclopedia of Clinical Psychology* (pp. 1–5). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118625392.wbecp549

Taxman, F. S. (2018). Risk Assessment: Where Do We Go from Here? In J. P. Singh, D. G. Kroner, J. S. Wormith, S. L. Desmarais, & Z. Hamilton (Eds.), *Handbook of Recidivism Risk / Needs Assessment Tools*. John Wiley & Sons, Incorporated.

Thornton, D., Mann, R., Webster, S., Blud, L., Travers, R., Friendship, C., & Erikson, M. (2003). Distinguishing and combining risks for sexual and violent recidivism. *Annals of the New York Academy of Sciences*, *989*, 225–235; discussion 236-246. https://doi.org/10.1111/j.1749-6632.2003.tb07308.x

Yang, M., Wong, S. C. P., & Coid, J. (2010). The efficacy of violence prediction: A meta-analytic comparison of nine risk assessment tools. *Psychological Bulletin*, *136*(5), 740–767. https://doi.org/10.1037/a0020473

Zhang, J., & Liu, N. (2015). Reliability and Validity of the Chinese Version of the LSI-R With Probationers. *International Journal of Offender Therapy and Comparative Criminology*, *59*(13), 1474–1486. https://doi.org/10.1177/0306624X14538396