Interrater Reliability of the LIJ: A Study to the Interrater Reliability of the National Instrument of the Juvenile Criminal Justice System (Summary)

L. Andries van der Ark, Julia L. van Leeuwen, Terrence D. Jorgensen

University of Amsterdam

Introduction (Chapter 1)

In this study, we investigated the *interrater reliability* (IRR) of the *Landelijk Instrumentarium Jeugdstrafrechtketen* (LIJ; National Instrument of the Juvenile Criminal Justice System). The LIJ instrument is used for flagging, screening, and risk assessment of each minor in The Netherlands who is suspect in a criminal case. Based on file research, interviews with the juvenile delinquent and the parents or care takers, and additional information, an officer of the *Raad voor de Kinderbescherming* (Child protection services) completes the LIJ. The IRR of the LIJ is the degree of agreement among different officers assessing the same juvenile delinquent by means of the LIJ. If the IRR is high, the officers will come to approximately the same assessment. If the IRR is low, the ratings of the officers differ substantially, and the outcome of the assessment thus depends on the officer who happened to be assigned to the case. Because the juvenile delinquent's punishment advice and possibly a plan of action are based on the LIJ assessment, the LIJ has a large impact on the juvenile delinquent, and a low IRR is undesirable. In this report, the IRR was estimated for both the dynamic risk profile and all individual LIJ items.

Technical chapters (Chapters 2, 3, and 4)

This study required research methodology that is not available in the standard literature, and had to be developed. The technical chapters describe the development of this methodology. Chapter 2 discusses the best possible estimation methods for IRR for complex assessments such as the LIJ. Chapter 3 describes a study into guidelines for the interpretation of IRR. The results enable researchers to answer questions such as "the reported IRR equals 0.37; what does that mean?". Chapter 4 describes a study investigating the best suitable research design for determining the IRR of the LIJ. An important result in Chapter 4 is that it is expected that approximately 150 completed LIJs are required for a fairly precise estimate of the IRR, but more than 500 are required for precise estimates. With the exception of Table 2, reading the technical chapters is not necessary for understanding the remainder of the research report.

The IRR of the LIJ (Chapter 5)

Method. The IRR of the LIJ was investigated using groups of 2-4 officers handling the same cases in Rotterdam, Haaglanden, Noord Nederland, Gelderland and Overijssel. The officers were instructed to work independently and not to talk to each other about the case. Each officer studied the file; each officer was present during the conversations with the juvenile delinquent and the parents, while one officer conducted the conversation and the others watched and listened from an observation room; each officer received additional information from informants; and each officer completed the LIJ based on all available information. This yielded 61 completed LIJs. Halfway through the research, the setup was slightly modified to collect as many LIJs as possible.

Result. The most important result is that the IRR of the LIJ is generally low. This applies to many of the items and to the dynamic risk profile. The main limitation of this research was the small sample size; the results were based on 61 assessments, while at least 150 completed LIJs and ideally 500 or more were desirable. As a result, for many questions and factors of the dynamic risk profile, the IRR has been estimated inaccurately, especially for selection instrument 2B.

Unclear questions (Chapter 6)

Because the IRR had been estimated inaccurately due to the low number of ratings, we decided to collect other information on the quality of the items that could possibly help improving the accuracy of the IRR. A survey was presented to 149 officers. The officers were asked to indicate which LIJ items (including answer options) they found unclear. 55 officers (37%) replied. 27 of the 131 items were perceived as unclear by more than 10% of the officers. Especially in Domains 8 (Attitude) and 10 (Vaardigheden [Skills]), a relatively large amount of items were found to be unclear. As expected, the correlation between the percentage of officers who found an item unclear and the estimated IRR of the item was negative (r = -. 12); and although statistically significant, the correlation was too weak to improve the accuracy of the IRRs.

Conclusion, discussion and recommendations (Chapter 7)

The IRR of the LIJ was generally low, while the IRR was also generally imprecisely estimated due to the small number of completed LIJs. Because all IRRs were estimated using the same small - and possibly unrepresentative - sample of completed LIJs, we do not have enough information to ascertain that the true IRRs are also low. For possible explanations of the small sample, we refer to the main text of Chapter 7.

Despite these limitations, all signs indicate that the IRRs of the LIJ can be improved. Based on discussions with officers and the survey, we see three possible causes of the low IRR that are relatively easy to remediate. First, we have noticed that the standardization of the administration of the LIJ is far from optimal: We found that the procedure for administering the LIJ differed across the different regions, and also across officers within a region. If a test-administration procedure is not standardized, the IRR tends to be low and the resulting test scores should not be compared across juvenile delinquents. Training and monitoring the test-administration procedure may help to increase the standardization, and therefore help to increase the IRR.

Second, we found that the time lag between the interviews with the juvenile delinquent and his or her parents until the officer filled out the LIJ was too long, sometimes more than a week. It is highly likely that the officer will forget the details and will base the assessment on global impressions only (halo effect). It is important that the time lag decreases, and that the LIJ is completed during the interviews, and possibly updated after new information has become available. If completing the LIJ during the interview is difficult or impossible, one could also consider using recordings of the interviews, which can serve as a reference when completing the LIJ.

Finally, the survey showed that a few items were unclear to the officers. The most unclear items may be revised using the outcomes of the survey in this report and the expertise of the officers.