Available online at www.sciencedirect.com

ScienceDirect

Procedia Social and Behavioral Sciences 5 (2010) 488–492

Procedia
Social and Behavioral Sciences

WCPCG-2010

Initial test of a new risk-need assessment instrument for youths with or at risk for conduct problems: ESTER-assessment

Henrik Andershed[a] *, Jennie Fredriksson[a], Karin Engelholm[a], Rickard Ahlberg[a], Steve Berggren[a], Anna-Karin Andershed[a]

[a]*Örebro University, School of Law, Psychology, and Social Work, Örebro, SE-701 82, Sweden*

Received January 2, 2010; revised February 3, 2010; accepted February 29, 2010

Abstract

ESTER-assessment is a new assessment instrument for youths (0-18 years), and includes 19 empirically-derived risk and protective factors for conduct problems. This study tests the inter-rater reliability of the five-point rating scale used to assess the 19 factors in ESTER-assessment on 30 institutionalized girls and their file information. Exact agreement between raters varied from 38 to 72 percent on the 19 individual factors, a result much better than chance. Intra-class correlations of the two independent raters on the majority of the 19 individual factors were fair to good. In conclusion, the results lend support to the inter-rater reliability of ESTER-assessment.
© 2010 Elsevier Ltd. Open access under CC BY-NC-ND license.

Keywords: Risk-need assessment, conduct problems, antisocial behavior, youth, inter-rater reliability.

1. Introduction

Youths with conduct problems is a group with relatively high risk for persistent problems, and the risk is higher the earlier the conduct problems starts (see e.g., Moffitt & Scott, 2008). For the sake of the individuals themselves and the people who are close to them, as well as the victims of crime and society as a whole, effective interventions are a crucial part of the collective social responsibility. However, for us to be able to tailor and implement effective interventions, individuals' different risks, needs and competencies need to be considered. The three principles of risk, need, and responsivity have been put forth as important parts of any successful intervention. This has been confirmed in empirical research showing that interventions that uphold these principles are more effective than interventions that do not (Andrews et al., 1990; Dowden & Andrews, 1999, 2002, 2003).

To be able to adhere to these principles, a risk-need assessment is needed and it is essential that this kind of assessment is reliable. One crucial type of reliability has to do with to what extent independent professionals make similar assessments of the same youth and his or her risk and protective factors. This so called inter-rater reliability refers to the degree of agreement between different raters of the same case. High inter-rater reliability means less subjectivity and improves legal security for the individual.

There are different assessment instruments used in practice today, such as the Structured Assessment for Violence Risk in Youth (SAVRY; Borum, Bartel, & Forth, 2002), Early Assessment Risk List for Boys (EARL-

* Seppo J. Laukka. Tel.: +358-8-5533658; fax: +358-8-5533600.
E-mail address: seppo.laukka@oulu.fi.

1877-0428 © 2010 Published by Elsevier Ltd. Open access under CC BY-NC-ND license.
doi:10.1016/j.sbspro.2010.07.129

20B; Augimeri, Webster, Koegl, & Levene, 1998) and Early Assessment Risk List for Girls (EARL-21G; Levene et al., 2001). SAVRY is a risk-need assessment instrument for youth 12-18 years, focusing on risk for future violence and covers risk as well as protective factors. The ratings of the risk factors are done on a three-point rating scale (Low, Moderate, High), while the protective factors are rated Present/Not present. EARL assesses boys (EARL-20B) and girls (EARL-21G) up to age 12 with conduct problems. EARL-20B covers 20 and EARL-21G, 21 risk factors, and uses a three-point rating scale (0, 1, 2) (Augimeri et al., 1998; Levene et al., 2001).

Most studies on the inter-rater reliability of SAVRY and EARL-20B conducted so far (no published studies have been found concerning the inter-rater reliability of EARL-21G) uses very small samples (*n*'s between 10 and 25) and almost exclusively present intra-class correlations (ICC's) on an aggregated level (i.e., several factors taken together) rather than on the individual factor level. These studies report ICC's between .52- .98 (e.g., Catchpole & Gretton, 2003; Dolan & Rennie, 2008; Enebrink et al., 2006; Lodewijks, Doreleijers, de Ruiter, & Borum, 2008; Lodewijks, Doreleijers, & de Ruiter, 2008; Meyers & Schmidt, 2008; Viljoen et al., 2008). Important to note is that the use of aggregated scales generally produces higher ICC's than use of individual factors.

ESTER-assessment (see Andershed & Andershed, 2008, 2010) is a new risk-need assessment instrument applicable to youths between 0 and 18 years of age. It includes 19 empirically-derived and practically relevant risk- and protective factors for conduct problems, which are grouped in four categories: nine Youth risk factors, three Family risk factors, four Youth protective factors, and three Family protective factors. ESTER-assessment can be used from first assessment/intake to case closure (i.e., is developed for multiple, repeated assessments), and to improve collaboration between professions. In ESTER-assessment, the rater decides what period of time that should be covered in the assessment, where a time-window between one and 36 months can be chosen.

ESTER-assessment uses a five-point rating scale to assess each of the 19 factors. The definitions of each scale-step for the youth and family *risk factors* are: *Not known* = The information is insufficient concerning all these behaviors during the period in question; *Not present (0)* = None of the above behaviors have been present during the period*; Weak (1)* = Does not occur often or is only causing very limited problems for the youth or his/her surroundings*; Evident (2)* = Occurs pretty often or is causing problems to some extent for the youth or his/her surroundings; *Pronounced (3)* = Occurs often or is causing extensive problems for the youth or his/her surroundings; *Very pronounced (4)* = Occurs very often or is causing extensive and serious problems for the youth or his/her surroundings. The definitions of each scale-step for the youth and family *protective factors* are: *Not known* = The information is insufficient concerning all these behaviors during the period in question; *Not present (0)* = None of the above behaviors have been present during the period*; Weak (1)* = Present but very limited in scope or is not at all strong or pronounced; *Evident (2)* = Present but not especially comprehensive, strong or pronounced; *Pronounced (3)* = Present and comprehensive, strong or pronounced; *Very pronounced (4)* = Present and very comprehensive, strong and pronounced (see Andershed & Andershed, 2010, for more details).

The purpose of this study is to test the inter-rater reliability of the five-point rating scale used to assess the 19 factors in ESTER-assessment. This was done based on file information of a sample of previously incarcerated girls.

2. Method

2.1. *Participants*

The file information of 30 girls previously incarcerated at an institution in Sweden for serious psychosocial problems, criminal behavior and/or drug abuse was used in the present study. A total of 42 files for girls meeting the inclusion criteria were found. It was possible to contact and ask for verbal consent for 33 of these, and of them 30 girls left their written consent that their file-information could be used in the study. The 30 girls ranged in age from 16 to 19 years and originated from various parts of Sweden. The criteria for inclusion in the sample were that the girl had stayed at the institution for at least eight weeks within the past five years, and had been subject to an eight-week evaluation procedure at the institution.

2.2. *Procedure*

Two independent ESTER-assessments were conducted based on file-information of the 30 girls. The assessments were conducted by two senior clinical psychology students and one graduate student in psychology, all with formal

training in ESTER. The raters did not communicate about the cases in question, prior or during the assessment. For the current study, a time-window of four months was chosen for the ESTER-assessment. Each ESTER-assessment took on average about two hours to complete. The files that were used to conduct the ESTER-assessments consisted of, for example; Adolescent Drug Abuse Diagnosis (ADAD; Friedman & Utada, 1989); Alcohol Drug Diagnosis Instrument (ADDIS; Hoffman et al., 1987); summaries of interviews with the youth and their parents covering psychosocial history, school, peer relations, etc.; psychological tests such as the Wechsler Intelligence Scale for Children (WISC-IV; Wechsler, 2003, 4th ed.) and the Wechsler Adult Intelligence Scale (WAIS-III; Weschler, 1999, 3rd ed.); pedagogical evaluation regarding concentration abilities, endurance, school knowledge; various file notes and information from social services of reason for placement at the institution; the institutions' observations of the girl, etc.

3. Results

Table 1 shows in the first three columns the agreement in frequency and percentage between the independent raters on the five-point rating scale on the 19 individual factors. The test of "Exact agreement" is the most stringent test of the inter-rater reliability. It tests to what extent the raters have assessed the factor in question exactly on the same rating step across the five-point rating scale. As seen in Table 1, exact agreement varied from 38 to 72 percent on the 19 factors. The test of "Exact agreement or difference in one step" varied from 77 to 100 percent (i.e., when for example rater A has assessed a factor as "3", then rater B has rated that factor as "2","3", or "4"). "Total disagreement" (i.e., rater A has rated the factor as "0" and rater B as "4" or vice versa) was non-existent in 16 out the 19 factors. It occurred on two factors and only on one out of 30 assessments on these two factors (see Table 1). To judge the intra-class correlations (ICC's) shown in the right hand column of Table 1, the following definitions were used; *poor* = < .40, *fair* = .40-.59, *good* = .60-.74, *excellent* = .75-1.00 (Chiccetti, 1994). As seen in Table 1, the ICC's were fair to excellent on 16 out of the 19 factors. On three factors, the ICC's were poor. However, as seen in the table, the absolute agreements on these three factors were not particularly low.

4. Discussion

How good are these levels of agreement across the two independent assessments? A comparison one could make is the one with chance, or random. We would definitely expect structured assessments as the ones conducted via ESTER-assessment to be in much greater agreement than an agreement gained by chance. If both assessments were totally random, the probability for the two random assessments of the same factor to be exactly the same on the 0-4 rating scale used in ESTER-assessment, would be four percent (based on the formula 1/5 x 1/5 = 0.04). Thus, we clearly see much better agreements than what would be achieved by chance between two independent ESTER-assessments. Exact agreement between raters varied from 38 to 72 percent on the 19 individual factors, figures much higher than four percent.

The exact agreement or difference in one step-agreement was quite high for all factors (77 to 100 percent). This is an important finding because a difference in one step on this five-point scale is not likely to generally be of critical clinical importance in the decision to intervene or not in the individual case. Another important finding, with high clinical significance, is the near total absence of total disagreement between raters.

Table 1. Inter-Rater Agreement on The ESTER-assessment Factors Between Two Independent Raters Measured Through Agreement In Frequency/Percent and Intra-Class Correlations (ICC).

	Exact agreement	Exact agreement or difference by one step	Total disagreement	ICC[a] (95% CI)
1.Defiant behavior, anger, or fearlessness	14/30 (47%)	26/30 (87%)	0/30 (0%)	.59*** (.29-.78)
2.Overactivity, impulsiveness or concentration difficulties	10/26 (38%)	20/26 (77%)	0/30 (0%)	.60*** (.28-.80)
3.Difficulties with empathy, feelings of guilt or remorse	11/26 (42%)	20/26 (77%)	0/26 (0%)	.66*** (37-.83)
4. Insufficient verbal abilities or school performance	14/30 (47%)	25/30 (83%)	1/30 (3%)	.53*** (.22-.75)
5.Negative problem solving, interpretations or attitudes	18/30 (60%)	28/30 (93%)	0/30 (0%)	.72*** (.49-.86)
6.Depressive mood or self harming behavior	13/30 (43%)	24/30 (80%)	1/30 (3%)	.59*** (.29-.78)
7.Conduct problems	16/30 (53%)	25/30 (83%)	0/30 (0%)	.60*** (.31-.79)
8. Alcohol or drug abuse	21/29 (72%)	29/29 (100%)	0/29 (0%)	.89*** (.77-.95)
9.Problematic peer relations	15/27 (56%)	25/27 (93%)	0/27 (0%)	.72*** (.47-.86)
Youth risk factors total				.78*** (.58-89)
10.Parents' own difficulties	13/27 (48%)	22/27 (82%)	0/27 (0%)	.77*** (.56-.89)
11.Difficulties in parent-youth relations	12/29 (41%)	25/29 (86%)	0/29 (0%)	.20 (-.18-.52)
12.Parents' difficulties with parenting strategies	13/29 (45%)	26/29 (90%)	0/29 (0%)	.49*** (.16-.72)
Family risk factors total				.62*** (.34-.80)
Youth and family risk factors total				.67*** (.41-.83)
13.Positive school attachment and performance	21/30 (70%)	26/30 (87%)	0/30 (0%)	.38* (.03-.65)
14.Positive attitudes and problem solving	11/28 (39%)	24/28 (86%)	0/28 (0%)	.55*** (.23-.76)
15.Positive relations and activities	15/26 (58%)	25/26 (96%)	0/26 (0%)	.64*** (.35-.82)
16.The youths' awareness and motivation	16/29 (55%)	25/29 (86%)	0/29 (0%)	.51** (.19-.74)
Youth protective factors total				.71*** (.48-.85)
17.Parents' energy, engagement and support	11/28 (39%)	24/28 (86%)	0/28 (0%)	.33* (-.04-.62)
18.Parents' positive attitudes and parenting strategies	10/23 (44%)	21/23 (91%)	0/23 (0%)	.53** (.28-.78)
19. Parents' awareness and motivation	12/30 (40%)	27/30 (90%)	0/30 (0%)	.58*** (.28-.78)
Family protective factors total				.68*** (.43-.84)
Youth and family protective factors total				.58*** (.28-.78)

Note. *p<.05; **p< .01; ***p<.001. [a] Single measure ICC. CI = Confidence Interval.

In comparison with other instruments like EARL-20B and SAVRY, the results of this study on ESTER-assessment seem to stand quite well. Studies on these other instruments report ICC's largely in line with the magnitude of the ICC's gained for the ESTER-assessment factors in the present study (e.g., Catchpole & Gretton, 2003; Dolan & Rennie, 2008; Enebrink et al., 2006; Lodewijks, Doreleijers, de Ruiter, & Borum, 2008; Lodewijks, Doreleijers, & de Ruiter, 2008; Meyers & Schmidt, 2008; Viljoen et al., 2008). This is interesting since both SAVRY and EARL-20B has fewer scale-steps than ESTER-assessment, which at least statistically improves the chance for raters to achieve higher inter-rater reliability.

Three of the ESTER-assessment factors had low/poor ICC's (i.e., lower than .40). These were factor *11. Difficulties in parent-youth relations,* factor *13: Positive school attachment and performance,* and factor *17: Parents' energy, engagement and support.* However, the exact agreement on these factors were quite high (see Table 1) and much higher than would be expected by chance. Thus, the low ICC's does not mean that these factors have low inter-rater reliability.

Some methodological limitations of the present study need mentioning. The sample used was quite small and only girls were included. Furthermore, file information was used as the single source of information for the ESTER-assessments. The ESTER-manual (Andershed & Andershed, 2008) specifies that at least two different kinds of sources or informants ideally should be used. Thus, future studies need to study the inter-rater reliability of ESTER-assessment using larger samples of boys and girls and using at least two different sources of information for the assessment. A central question for future studies is also whether one reaches higher inter-rater reliability with ESTER-assessment than when conducting the assessment in an unstructured way, without an instrument.

5. Conclusion

Inter-rater reliability is an essential feature of risk-need instruments. The results of the present study lend support to the inter-rater reliability of ESTER-assessment.

References

Andershed, H., & Andershed, A-K., (2008). *The ESTER-manual: Structured assessment and follow-up of research-based risk and protective factors in youths with or at risk for conduct problems*.

Andershed, H., & Andershed, A-K. (2010). Risk-need assessment for youth with or at risk for conduct problems: Introducing the assessment system ESTER. *Procedia Social and Behavioral Journal.*

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

Augimeri, L. K., Webster, C. D, Koegl, C. J., & Levene, K. S. (1998). *Early Assessment Risk List for Boys: EARL-20B. Version 1: Consultation edition.* Toronto, Canada: Earlscourt Child and Family Centre.

Borum, R., Bartel, P., & Forth, A. (2002). *Manual for the Structured Assessment for Violence Risk in Youth (SAVRY), Consultation edition, Version 1.* Tampa: University of South Florida.

Catchpole, R., & Gretton, H. (2003). The predictive validity of risk assessment with violent young offenders: A 1-year examination of criminal outcome. *Criminal Justice & Behavior, 30*, 688-708.

Chiccetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6,* 284-290.

Dolan, M. C., & Rennie, C. E. (2008). The Strucutred Assessment of Violence Risk in Youth as a predictor of recidivism in a United Kingdom cohort of adolescent offenders with conduct disorders. *Psychological Assessment*, 20, 35-46.

Dowden, C., & Andrew, D. A. (1999). What works in young offender treatment: A metaanalysis. *Forum on Corrections Research, 11*, 21-24.

Dowden, C., & Andrews, D. A. (2002). A meta-analytic examination of the principles of effective correction interventions for young female offenders. I A. Cummings & A. Leschied (Red.), *Research and treatment for aggression with adolescent girls* (page. 133-160). Lewiston, NY: The Edwin Mellen Press.

Dowden, C., & Andrews, D. A. (2003). Does family intervention work for delinquents? Results of a meta-analysis. *Canadian Journal of Criminology and Criminal Justice, 45*, 327-342.

Enebrink, P., Långström, N., Hultén, A., & Gumpert, C. H. (2006). Swedish validation of the Early Assessment Risk List for Boys (EARL-B), a decision aid for use with children presenting with conduct disordered behaviour. *Nordic Journal of Psychiatry, 60,* 438-446.

Friedman, A. S., & Utada, A. (1989). A method for diagnosing and planning the treatment of adolescent drug abusers: The Adolescent Drug Abuse Diagnosis (ADAD) instrument. *Journal of Drug Education, 19*, 285-312.

Hoffman, N., Normann, G., Harrison, P. A., & Wickström, L. (1987). *ADDIS – Alkohol, Drog Diagnos Instrument, manual*. Åre: 4M Konsult AB.

Levene, K. S., Augimeri, L. K., Pepler, D. J., Walsh, M. M, Webster, C. D., & Koegl, C. J. (2001). *Early Assessment Risk List for Girls (EARL-21G).* Version 1 - Consultation Edition. Toronto, Canada: Earlscourt Child and Family Centre.

Lodewijks, H. P. B., Doreleijers, T. A.H., de Ruiter, C., & Borum, R. (2008). Predictive validity of the Structured Assessment of Violence Risk in Youth (SAVRY) during residential treatment. *International Journal of Law and Pyschiatry*, 31, 263-271.

Lodewijks, H., Doreleijers, T. A. H., de Ruiter, C. (2008). SAVRY Risk Assessment in Violent Dutch Adolescents. Relation to Sentencing and Recidivsm. *Criminal Justice and Behavior, 35*, 696-709.

Meyers, J., R., Schmidt, F. (2008). Predicitve Validity of the Structured Assessment for Violence Risk in Youth (SAVRY) with Juvenile Offenders. *Criminal Justice and Behavior, 35*, 344-355.

Moffitt, T. E., & Scott, S. (2008). Conduct disorders of childhood and adolescence. In M. Rutter, D. Bishop, D. Pine, S. Scott, J. Stevenson, E. Taylor, & A. Thapar (Eds.), *Rutter´s Child and Adolescent Psychiatry,* (5th ed) (pp. 543-564). Oxford: Blackwell Publishing.

Viljoen, J. L., Scalora, M., Cuadra, L., Bader, S., Chávez, V., Ullman, D., & Lawrence, L. (2008). Assessing risk for violence in adolescents who have sexually offended. A Comparison of the J-SOAP-II, J-SORRAT-II, and SAVRY. *Criminal Justice and Behavior, 35*, 5-23.

Wechsler, D. (2003). *The Wechsler Intelligence Scale for Children (4rd ed.).* San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1999). *The Weschler Adult Intelligence Scale (3rd ed.).* San Antonio, TX: The Psychological Corporation.