American Journal of Applied Psychology
2013; 2(2) : 16-21
Published online March 10, 2013 (http://www.sciencepublishinggroup.com/j/ajap)
doi: 10.11648/j. ajap.20130202.11

Further testing of the inter-rater reliability of ESTER-assessment - a risk-need assessment instrument for youths with or at risk for conduct problems

Henrique Bond, Marja Rudenhed, Eva Bergquist, Anna-Karin Andershed, Henrik Andershed*

School of Law, Psychology, and Social Work, Örebro University, SE-701 82, Sweden

Email address:
henrik.andershed@oru.se (H. Andershed)

To cite this article:
Henrique Bond, Marja Rudenhed, Eva Bergquist, Anna-Karin Andershed, Henrik Andershed. Further Testing of the Inter-rater Reliability of ESTER-assessment – a Risk-need Assessment Instrument for Youths with or at Risk for Conduct Problems, *American Journal of Applied Psychology*. Vol. 2, No. 2, 2013, pp. 16-21. doi: 10.11648/j.ajap.20130202.11

Abstract: Behavioral problems in childhood have been associated with conduct problems later in life. Thus, it is essential that youths with or at risk for conduct problems receive the help they need on time. Therefore, youth with or at risk for conduct problem must receive effective risk-need assessments and intervention plans regardless the person who conducts the assessment. ESTER-assessment is a structured, computer-aided, risk-need instrument developed for assessing youth (0-18) with or at risk for conduct problems. It uses a five-step response scale to assess 19 research-based risk and protective factors and the present study tests the inter-rater reliability of these 19 factors. This was done by comparing the assessments conducted by two independent raters who assessed the file information of 30 girls (mean age = 16.9) who had been incarcerated due to psychosocial problems, criminality and/or drug abuse. Results showed fair to good agreement for the majority of the factors via intra-class correlations and percentage agreement varied on the 19 factors from 24.1 to 80.8 % for exact agreement and from 72.2 to 96.7 % for exact agreement or difference by one step on the response scale. We conclude that it is possible to gain acceptable to excellent inter-rater reliability in assessing risk and protective factors via ESTER-assessment.

Keywords: Risk-Need Assessment, Conduct Problems, Antisocial Behavior, Youth, Inter-Rater Reliability

Early conduct problems have been associated with increased risk for negative psychosocial development [1, 2, 3] and research shows that the earlier the start of problem behavior, the greater the risk for negative development [4, 5]. Therefore, it is essential that professionals intervene early and have access to reliable assessments tools that can assist in identifying the youth's risks and needs at an early stage, thereby facilitating the implementation of effective interventions. One central aspect of reliability is inter-rater reliability, that is, the extent to which independent raters agree when assessing something. The purpose of the present study is to test the inter-rater reliability of ESTER-assessment - Evidence-based STructured assEssment instrument of Risk and protective factors [6].

An increasing number of empirical studies suggest that three principles are essential for enhancing the efficacy of assessments and thereby facilitating the choice of adequate interventions: Risk, need, and responsivity [7, 8]. The risk principle points out that the degree of risk of the individual should guide the extensiveness and intensiveness of the intervention. Extensive and intensive interventions are most effective for high risk youths. The need principle concerns what should be targeted in assessments and interventions, and states that the focus should be on identifying the most relevant and changeable, criminogenic factors (i.e., factors that are changeable and directly associated with the youth's criminality/conduct problems). The responsivity principle concerns how the intervention should be carried out in order to gain responsivity in the youth and proposes that interventions should match the individual characteristics of the youth such as learning style and motivation [9, 10]. A research review [10] has shown that interventions based on these three principles are more effective than interventions that are not. Hence, it is important that an assessment instrument helps the professional to assess the degree of risk as well as individual patterns of risk and protective factors, so that the risk, need, and responsivity principles can be adhered to.

Three assessment instruments that are used today on youths with conduct problems are the Structured Assessment for Violence Risk in Youth (SAVRY) [11], for youths between 12 and 18-years old, the Early Assessment Risk List for Boys under age 12 (EARL-20B) [12] and the Early Assessment Risk List for Girls under age 12 (EARL-21G) [13]. Several studies have tested the inter-rater agreement/reliability of these instruments but most of them using very small samples (*n*'s ranging from 10 and upward) and have almost exclusively focused on results from Intra-Class Correlations (ICC's) and on an aggregated level (i.e., several factors/items taken together) rather than on the individual factor/item level. These studies report ICC's between .34 and .84 for SAVRY [14] and .17 and .71 on EARL-20B [15] on the individual factor/item level. On the aggregated level the ICC's varied between .81 and .97 for SAVRY [16, 17, 18, 19, 20, 21], .64 and .82 on EARL-20B [15, 22] and between .80 and .86 on EARL-21G [23]. Important to note is that the use of aggregated scales generally produces higher ICC's than use of individual factors/items.

In terms of percentage agreement, exact agreements vary between 55.0 and 93.0 % on the factors/items of EARL-20B [18] and between 35.0 and 94.0 % on the factors/items of SAVRY [14]. With regard to total disagreement between raters (i.e., one rater assess the factor as lowest possible and the other rater as highest possible), the only study reporting this found that this happened on two of the 20 EARL-20B factors in 3.3 % of the ratings on the factor "Onset of Behavioral Difficulties" and in 16.6 % of the ratings on the factor "Abuse/Neglect/Trauma" [18].

ESTER-assessment is an evidence-based structured computer-aided assessment instrument developed for children and adolescents between 0 and 18 years of age with or at risk for conduct problems [6]. ESTER-assessment focuses on 19 dynamic (i.e., potentially changeable) risk- and protective factors grouped in four categories: Youth risk factors, Family risk factors, Youth protective factors, and Family protective factors (see Table 1). ESTER-assessment is different from SAVRY and EARL-20B and 21G because (1) it applies to both children and adolescents, (2) it can be used for both boys and girls , (3) it uses a five-step rating scale that enables the detection of quite small but important changes over time (e.g., during or after an intervention), which improves the possibility to use the instrument for continuous follow-ups and evaluations; and (4) it is supported by a computerized system that facilitates interpretation of the assessments [6].

Table 1. *Inter-rater Reliability of the ESTER-assessment Factors Tested Through Agreement in Percent Compared to Agreement by Chance and Through Intra-class Correlation Coefficients (ICC).*

	Exact agreement[a]	**Exact agreement or difference by one step**[b]	**Total disagreement**	**ICC**[c] **(95% CI)**
Youth Risk Factors				
1. Defiant behavior, anger, or fearlessness	41.4% (12/29)**	86.2% (25/29)**	0.0% (0/29)	.17 (-.20-.50)
2. Overactivity, impulsiveness or concentration difficulties	30.8% (8/26)*	88.5% (23/26)**	0.0% (0/26)	.83 (.66-.92)***
3. Difficulties with empathy, feelings of guilt or remorse	38.9% (7/18)*	72.2% (13/18)**	16.7% (3/18)	.50 (.06-.78)*
4. Insufficient verbal abilities or school performance	33.3% (6/18)*	83.3% (15/18)**	0.0% (0/18)	.76 (.47-.90)***
5. Negative problem solving, interpretations or attitudes	28.6% (6/21)*	90.5% (19/21)**	0.0% (0/21)	.69 (.38-.86)***
6. Depressive mood or self harming behavior	37.0% (10/27)**	88.9% (24/27)**	3.7% (1/27)	.58 (.27-.79)**
7. Conduct problems	70.0% (21/30)**	93.3% (28/30)**	0.0% (0/30)	-.10 (-.44-.27)
8. Alcohol or drug abuse	80.8% (21/26)**	96.2% (25/26)**	0.0% (0/26)	.72 (.47-.86)***
9. Problematic peer relations	41.7% (10/24)**	75.0% (18/24)**	0.0% (0/24)	N.A.[1]
Family Risk Factors				
10. Parents' own difficulties	66.7% (14/21)**	95.2% (20/21)**	0.0% (0/21)	.91 (.79-.96)***
11. Difficulties in parent-youth relations	29.6% (8/27)**	88.9% (24/27)**	0.0% (0/27)	.37 (-.00-.66)*
12. Parents' difficulties with parenting strategies	66.7% (14/21)**	92.2% (20/21)**	0.0% (0/21)	N.A.[1]
Risk Factors Total				.38(-.08-.71)*
Youth Protective Factors				
13. Positive school attachment and performance	35.0% (7/20)*	85.0% (17/20)**	0.0% (0/20)	.42 (-.01-.72)*
14. Positive attitudes and problem solving	46.7% (14/30)**	96.7% (29/30)**	0.0% (0/30)	.61 (.33-.80)***
15. Positive relations and activities	29.7% (8/27)**	88.9% (24/27)**	0.0% (0/27)	.38 (.00-.66)*
16.The youth's awareness and motivation	70.0% (21/30)**	93.3% (28/30)**	0.0% (0/30)	.47 (.13-.70)**
Family Protective factors				
17.Parents' energy, engagement and support	24.1% (7/29)*	86.2% (25/29)**	0.0% (0/29)	.31 (-.05-.61)*
18.Parents' positive attitudes and parenting strategies	64.3% (9/14)**	92.9% (13/14)**	0.0% (0/14)	.43 (-.11-.77)
19. Parents' awareness and motivation	37.9% (11/29)**	79.3% (23/29)**	0.0% (0/29)	.47 (.13-.71)**
Protective Factors Total				.37 (.01-.64)*

Note. *p < .05; **p < .01; ***p < .001. [a]Percentages of agreement are tested against the percentage of agreement by chance via z-tests (one-tailed). The observed percentages of Exact agreement (e.g., when both rater A and B assessed the same factor as "4") are tested against 4 percent (agreement by chance according to the formula: 1/5 x 1/5 = 0.04). [b]The observed percentages of exact agreement or difference in one step (e.g., when rater A assessed a factor as "3" and rater B assessed the same factor as "2", "3" or "4") are tested against 12 percent (agreement by chance according to the formula: 1/5 x 3 = 0.12). [c]Single measure ICC. [1]Not analyzable due to lack of variation. CI = Confidence Interval.*

An ESTER-assessment is based on a predetermined period back in time – decided by the professional – where the time-window back in time is somewhere between 1 and 36 months. Each of the 19 risk and protective factors are explicitly defined and each factor include several concrete descriptions of behaviors or characteristics related to the definition of the factor, and each of the 19 factors are then rated on a five-point scale. The risk factors are rated using a scale ranging from Not present (0) to Very pronounced (4) with definitions for each scale step focusing on how frequent and severe/problematic the behaviors are. Each scale step is clearly defined. Similarly, protective factors are rated using a five-step scale but the definitions of each scale step are focused on how pronounced and comprehensive the protective factor is [6]. A manual called the ESTER-manual [6] provide a support for the professional conducting an ESTER-assessment and it specifies six core assessment principles that should be followed by the rater. For instance, one of the six principles specifies that when the behaviors of a specific factor are present in one context (e.g., school) but not in another (e.g., home), the raters are instructed to assess according to the information indicating the most frequent or problematic behavior when it comes to risk factors, and the weakest protection when it comes to protective factors [6].

An initial test of the inter-reliability of the five-step scale of ESTER-assessment has previously been conducted [24]. This previous study was based on ESTER-assessments conducted by two independent raters using file-information of 30 girls, between 16 and 19-years old, incarcerated for serious psychosocial problems, criminal behavior or drug abuse. The Intra-Class Correlations (ICC's) ranged between .49 and .89 on 16 out of the 19 individual factors. On the remaining three factors, the ICC-scores ranged between .20 and .38. Moreover, the aggregated ICC was .67 for the risk factors total and .58 for the protective factors total. The percentage agreement analyses were more convincing than the ICC's in this study. The exact percentage agreement between raters (i.e., when both raters A and B assess a factor exactly the same, for example as "Very pronounced") varied between 38.0 and 72.0 % on the 19 individual factors. Percentage scores of exact agreement or difference in one scale step (e.g., rater A assesses a factor as "3", and rater B assesses the same factor as "2", "3" or "4") varied between 77.0 and 100 %. Total disagreement (i.e., when rater A assesses a factor as "0" and rater B assesses the same factor as "4", or vice versa) was present in only two of the 19 factors one time each (Factor 4 "Insufficient verbal abilities or school performance" and Factor 6 "Depressive mood or self harming behavior"). In conclusion, this first study showed acceptable inter-rater reliability and close to as good percentage agreement on individual factors as the studies on SAVRY and EARL-20B and 21G [24]. This is quite striking because SAVRY uses three- and two-step scales on its factors and the EARL instruments a three-step scale. ESTER-assessment uses a five-point scale, making it more difficult to gain exact agreements.

The present study aims to test the inter-rater reliability of ESTER-assessment under similar conditions as the initial study [24] but with other raters. The present study partly uses the same sample as in the initial study but two other raters with the main question being whether the inter-rater reliability of ESTER-assessment can be generalized to other raters. The present study also tests whether the percentage agreements gained on the 19 factors between the two raters are significantly different for what would be gained by chance. This was not done in the initial study.

1. Method

1.1. Subjects

The ESTER-assessments were conducted on file-information of 30 girls who had been incarcerated in an institution for youths in Sweden. Twenty-two of those girls participated in the initial study on the inter-rater reliability of ESTER-assessment [24]. Eight of the girls' file information were not possible to access for the present study. Therefore, eight other girls were included for the present study. The reasons for incarceration were the youths' psychosocial problems, criminality and/or drug abuse. The age of the participants ranged between 15 and 20 years (Mean age = 16.9 years). The girls came from different regions in Sweden.

1.2. Procedure

Two independent raters, both formally trained in ESTER-assessment, conducted the assessments based on file information only. The raters did not discuss the assessments before or during the assessment process. The present study used a six month time-window for the ESTER-assessments. That is, the risk and protective factors were assessed based on file information from the present and six months back in time. The files used in this study consisted of information from different sources such as interviews (e.g., Adolescent Drug Abuse Diagnosis, ADAD) [25], documentation of offence history, psychological tests (e.g., Wechsler Intelligence Scale for Children - fourth edition, WISC-IV) [26] as well as pedagogical and social evaluations of the youth (e.g., file-summaries of interviews with the youth and parents or caregivers regarding psychosocial history, school, peer relations and socio-economic status). The vast majority of the 30 youths had already left the institution at the time this study was conducted. They were informed in writing and gave their written and active consent that their files could be used for the present research purpose.

1.3. Statistical Analyses

Inter-rater reliability was tested by comparing the two independent raters' assessments of the 19 risk and protective factors in ESTER-assessment. Agreement was studied both in terms of various percentage comparisons as well as with intra-class correlations.

2. Results

Exact agreement on the individual factor level can be seen as the strictest test of agreement (i.e., the extent both raters assess a factor exactly the same on the five-step scale, for example as "Very pronounced"). By chance, two independent raters would achieve exact agreement on an individual factor in 4.0 % of the assessments, based on the formula 1/5 x 1/5 = 0.04. As seen in Table 1, the exact agreement varied between 26.6 and 80.8 % for the risk factors in ESTER-assessment, and between 24.1 and 70.0 % for the protective factors. As shown in Table 1, these degrees of agreement were significantly higher (as shown via z-tests) than what would be expected by chance (i.e., 4.0 %), for all 19 factors.

With regard to exact agreement or difference by one step (e.g., when rater A has assessed a factor as "3" and rater B has assessed that factor as "2", "3" or "4" or vice versa), two independent raters would by chance achieve exact agreement or difference in one step on an individual factor in 12.0 % of the assessments, based on the formula 1/5 x 3 = 0.12. As shown in Table 1, exact agreement or difference by one step varied between 72.3 and 96.2 % on the risk factors and between 79.3 to 96.7 % on the protective factors. These agreement scores were significantly higher on all 19 factors than those gained by chance.

As seen in Table 1, total disagreements (i.e., when rater A assessed a factor as "0" and rater B assessed the same factor as "4" or vice versa) occurred only for two of the 19 factors. The two factors were Factor 3 and Factor 6 and the total disagreements happened in 16.7 and 3.7 % of the assessments, respectively.

Intra-Class Correlations (ICC's) were calculated for each individual risk and protective factor as seen in Table 1. An ICC of less than .40 is here considered as poor, values between .40 and .59 as fair, .60 to .74 as good, and .75 to 1.00 as excellent [27]. As seen in Table 1, the aggregated ICC concerning the risk factors total shows poor agreement (ICC = .38). Similarly, the aggregated ICC for the protective factors shows poor agreement (ICC = .37). Five individual factors also exhibit poor ICC's, varying between -.10 and .38. However, important to note is that the percentage agreement in terms of Exact agreement and Exact agreement or difference in one step on these individual factors with poor ICC's are quite high and significantly higher than would be expected by chance. Twelve of the 19 factors present fair to excellent ICC's, varying between .42 and .91. For two risk factors; Factors 9 and 12, it was not possible to calculate ICC's due to lack of variation. However, as seen in Table 1, those factors had percentage scores that were significantly higher than would be expected by chance concerning Exact agreement and Exact agreement or difference by one step.

3. Discussion

To provide effective help characterized by professionalism and legal security to youth with or at risk for conduct problems, assessment instruments with acceptable reliability are essential. The aim of this study was to investigate the inter-rater reliability of the five-step scale used in ESTER-assessment. The intra-class correlations indicate poor to excellent reliability of the 19 individual factors. However, the most strict test of inter-rater reliability; Exact agreement between the two raters on the five-step scale, shows that all 19 factors are significantly higher in agreement percentage than would be expected by chance. In addition, total disagreements between the two independent raters were very uncommon, something that should not be underestimated in importance.

The findings of the present study are to a large extent consistent with the results of the previous study on the inter-rater reliability of ESTER-assessment [24] where ICC's were fair to good and the percentage scores were generally quite similar as in the present study. In comparison with other instruments, such as EARL-20B and 21G and SAVRY, the inter-rater reliability of ESTER-assessment presents comparable degrees of agreement, with a tendency toward a somewhat lower agreement. However, such a comparison is problematic since the inter-rater reliability research on EARL and SAVRY almost exclusively have focused on aggregated scores instead of scores on individual factors [16, 17, 19, 22]. Aggregated ICC analyses may well hide very low individual factor ICC's. In terms of percentage agreements on individual factors, the present study's numbers varies between 24.1 and 80.8 % for Exact agreement. These results are lower than those for EARL-20B (between 55 and 93.0 %) [18] and SAVRY (between 35 and 94.0%) [14]. However, it should be pointed out that ESTER-assessment uses a five-step rating scale whereas the other two instruments use a two- or three-step scale. Thus, one could expect the agreement to be higher for the SAVRY and EARL instruments. The percentage results of exact agreement or difference by one step were quite high, with percentage scores varying from 72.3 to 96.7 % in the present study. Since ESTER-assessment uses a five-step scale, difference in one step should generally not lead to significant differences when it comes to choosing a matching intervention plan.

Some methodological limitations of the present study need attention. First, the study uses a small sample size (n = 30) and includes only girls. Future research should assess larger samples and include both girls and boys. Second, we used only one type of information (i.e., file information) to conduct the ESTER-assessments. This is not in line with the standard recommendations for conducting an ESTER-assessment, which states that one ideally should use at least two different sources or informants (e.g., interviewing a parent and a teacher and using file information if it exists). This deviation from the standard procedure may have led both to an under- or overestimation of the inter-rater reliability.

One of the strengths of the present study is that it includes percentage comparisons of raters' agreement as a complement to the intra-class correlations. This is an important aspect of agreement analyses because ICC-results can be misleading. For example, as we observed for Factor 7: Conduct problems, the ICC is very poor (ICC = -.10). In contrast, the percentage score for exact agreement; 70.0 %, is quite high. Another advantage of using percentage comparisons is that it allows illustrating degrees of agreements between raters even when variance is absent (i.e., when ICC calculations are not possible). For instance, we could not calculate ICCs for Factor 9: Problematic peer relations or Factor 12: Parents' difficulties with parenting strategies. However, the percentage scores show that Exact agreement between raters were higher than 40.0 % on both of these factors. In line with this reasoning, statistical expertise has argued that correlation coefficients (e.g., ICC) are not the most appropriate measure to assess levels of agreement, instead, percentage agreement should be used [28].

The inter-rater reliability gained with ESTER-assessment is not perfect and perhaps this is to be expected. A key question is whether this none-perfect reliability is higher than when an instrument is not used. This could for example be tested in a case vignette study where one group of professionals trained in ESTER-assessment would assess the same case, and another group of professionals not trained in ESTER-assessment would assess the same case. The hypothesis would be that the professionals using ESTER-assessment would identify more risk and protective factors in the vignette and to a greater extent agree in their ratings as compared to the professionals not using an instrument.

4. Conclusion

We conclude that it is possible to gain acceptable to excellent inter-rater reliability in terms of agreement in percentage in assessing risk and protective factors via ESTER-assessment. We replicate the previous study on ESTER-assessment using an overlapping sample (22 of the participants were the same) showing that the inter-rater reliability of ESTER-assessment can be generalized to other raters (i.e., the inter-rater reliability is due to the instrument rather than the raters). These findings show that, by using ESTER-assessment, raters can achieve acceptable inter-rater reliability in assessing individual risk- and protective factors for conduct problems in youth. Thus, using ESTER-assessment increases the chances for youths with or at risk for conduct problem to receive intervention plans that matches their needs regardless the person who conducts the assessment.

5. Funding

This project was supported by the Swedish National Board of Health and Welfare and the Swedish National Research Council. During the preparation of this manuscript Henrik Andershed was supported by funds from the Swedish National Research Council.

References

[1] Frick, P. J., & Viding, E. (2009). Antisocial Behavior from a Developmental psychopathology Perspective. Development and Psychopathology, 21, 1111-1131.

[2] Krohn, M. D., Thornberry, T. P., Rivera, C., & Le Blanc, M. (2001). Later delinquency careers. In R. Loeber & D. P. Farrington (Eds.), Child delinquents (pp. 67-94). Thousand Oaks, CA: Sage.

[3] Snyder, H. N. (2001). Epidemiology of Official Offending. In R. Loeber & D. P. Farrington (Red.), Child delinquents (pp. 25-46). Thousand Oaks, CA: Sage.

[4] 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.

[5] Odgers, C., Caspi, A., Broadbent, J. M., Dickson, N., Hancox, R. J., & Harrington, H. (2007). Prediction of Differential Adult Health Burden by Conduct Problem Subtypes in Males. Archives of General Psychiatry, 64, 476-484.

[6] 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, 5, 377-383.

[7] Andrews, D. A., Bonta, J., & Wormith, J. S. (2006). The Recent Past and Near Future of Risk and/or Need Assessment. Crime & Delinquency, 52, 7-27.

[8] 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.

[9] Andrews, D. A., Bonta, J., & Hoge, R. D. (1990). Classification for effective rehabilitation- Rediscovering Psychology. Criminal Justice and Behavior, 17, 19-52.

[10] Andrews, D. A., & Bonta, J. (2010). Rehabilitating criminal justice policy and practice. Psychology, Public Policy, and Law, 16, 39-55.

[11] 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.

[12] 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.

[13] 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.

[14] Vincent, G. M., Guy, L. S., Fusco, S. L., & Gerhenson, B. G. (2011). Field Reliability of the SAVRY with Juvenile Probation Officers: Implications for Training. Law and Human Behavior.doi: 10.1007/s10979-011-9284-2.

[15] Hrynkiw-Augimeri, L. K. (2005). Aggressive and antisocial young children: Risk prediction, assessment and management utilizing the Early Assessment Risk List for Boys (EARL-20B). Unpublished doctoral dissertation, Ontario Institute for Studies in Education, University of Toronto, Ontario, Canada.

[16] 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.

[17] Dolan, M. C., & Rennie, C. E. (2008). The Structured Assessment of Violence Risk in Youth as a Predictor of Recidivism in a United Kingdom Cohort of Adolescent Offenders With Conduct Disorder. Psychological Assessment, 20, 35-46.

[18] Enebrink, P., Långström, N., Hultén, A., & Gumpert, C. H. (2006). Swedish Validation of the Early Assessment Risk List for Boys (EARL 20B), a decision aid for use with children presenting with conduct-disrordered behaviour. Nordic Journal of Psychiatry, 60, 438-446.

[19] 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.

[20] Meyers, J., & Schmidt, F. (2008). Predictive Validity of the Structured Assessment for Violence Risk in Youth (SAVRY) with Juvenile Offenders. Criminal Justice and Behavior 35, 696-709.

[21] 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.

[22] Hrynkiw-Augimeri, L. K. (1998). Assessing risk for violence in boys: A preliminary risk assessment study using the Early Assessment Risk List for Boys (EARL-20B). Unpublished master's thesis, Ontario Institute for Studies in Education, University of Toronto, Ontario, Canada.

[23] Levene, K. S., Walsh, M. M., Augimeri, L. K., & Pepler, D. J. (2004). Linking Identification and Treatment of Early Risk Factors for Female Delinquency. In M. M, Moretti, C. L., Odgers, & M. A., Jackson (Eds.), Girls and Agression: Contibuting Factors and Intervention Principles (pp. 147-163). New York: Kluwer Academic/Plenum Publishers.

[24] Andershed, H., Fredriksson, J., Engelholm, K., Ahlberg, R., Berggren, S., & Andershed, A-K. (2010). Initial test of the new risk-need assessment instrument for youths with or at risk for conduct problems: ESTER-assessment. Procedia Social and Behavioral Sciences, 5, 488-492.

[25] 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.

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

[27] Cicchetti, D. V. (1994). Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment instruments in Psychology. Psychological Assessment, 6, 284-290.

[28] Svensson, E. (2001). Guidelines to Statistical Evaluation of Data from Rating Scales and Questionnaires. Journal of Rehabilitation Medicine, 33, 47-48.