Research in Developmental Disabilities 119 (2021) 104111

Contents lists available at ScienceDirect

Research in Developmental Disabilities

journal homepage: www.elsevier.com/locate/redevdis

Evaluating implementation outcomes of a measure of social vulnerability in adults with intellectual disabilities

Mireille Tabin [*], Cindy Diacquenod, Geneviève Petitpierre

Department of Special Education, University of Fribourg, Fribourg, Switzerland

ARTICLE INFO

The paper is part of a special issue on “Research Practices”.

Keywords:
Intellectual disabilities
Social vulnerability
Implementation outcomes
Assessment fidelity
Acceptability
Appropriateness
User experience

ABSTRACT

Background: A test identified as valid and accurate in research will not automatically be considered appropriate by those involved in its use, or even be used in the first place. The Social Vulnerability Test-22 items [TV-22] is a measure specially designed for adults with intellectual disabilities (ID). This study aims to evaluate the implementation outcomes of the TV-22; more precisely its acceptability (e.g., complexity), appropriateness (e.g., perceived relevance) and the assessment fidelity (i.e., adherence to assessment guidelines) by special education practitioners.
Procedures: Thirty-one practitioners (8 psychologists, 11 educators, 12 special education center managers) administered the TV-22 during an interview with an adult with ID. Semi-structured interviews were conducted to collect practitioners’ opinions on the acceptability and the appropriateness of the TV-22 for their clinical practice. Quantitative analyses were performed to assess the fidelity of the assessments and the influence of some personal factors.
Results: The results indicate a good appropriateness, a reasonable acceptability, – but a low assessment fidelity of the TV-22 by some practitioners. Psychologists stand out for a more rigorous use of the test.
Implications: Results highlight the importance of evaluating implementation outcomes when a new measure is developed to ensure its appropriateness and correct use by stakeholders.

What does this paper add?

The assessment of social vulnerability aims to evaluate the ability of a person to detect and handle potentially harmful social situations. The Social Vulnerability Test-22 items [TV-22] is a measure specially designed for adults with ID. Not all social institutions providing services for adults with ID have a psychologist on staff: the assessments are sometimes carried out by educational staff or the team/center manager. In this context therefore, evaluating to what extent the characteristics of the assessor are likely to influence the quality of the assessment is particularly important. In the present research, we aim to evaluate the implementation outcomes (acceptability, appropriateness, assessment fidelity) of the TV-22 by special education practitioners (psychologists, educators, team/center managers). The results indicate a good appropriateness, a reasonable acceptability, but a low assessment fidelity of the TV-22 by some practitioners – which definitely questions the quality of the social vulnerability assessment performed. Despite its important role in interpreting results, assessment fidelity has received little attention to date; more research is needed to explore assessment fidelity and possibilities to enhance it when required.

[*] Corresponding author.
E-mail address: mireille.tabin@unifr.ch (M. Tabin).

https://doi.org/10.1016/j.ridd.2021.104111
Received 30 April 2021; Received in revised form 30 September 2021; Accepted 5 October 2021
Available online 9 October 2021

1. Introduction

A test considered as valid and accurate in research will not automatically be considered appropriate by those involved in its use, or even be used in the first place. Greenhalgh et al. (2004) discuss the need to move from a "let it happen" mindset to a "help it happen" or even "make it happen" perspective. Evaluating implementation outcomes is one possibility to make this move into the "make it happen" perspective. Implementation outcomes refer to "effects of deliberate and purposive actions to implement new treatments, practices, and services" (Proctor et al., 2011, p. 65). Evaluating implementation outcomes is important: implementation outcomes are not only indicators of the implementation success and processes, they are also essential features of program or innovation evaluations as they affect the outcomes obtained (Durlak & DuPre, 2008).

Acceptability, appropriateness, and fidelity, are part of the core set of implementation outcomes (Proctor et al., 2011). Acceptability refers to the satisfaction of implementation stakeholders with various aspects of the innovation (e.g., content, complexity, credibility). Appropriateness is the perceived relevance, suitability and compatibility of the innovation for a given practice setting. Whereas "appropriateness" and "acceptability" are conceptually similar and sometimes overlapping, it seems accurate to distinguish between them because a given innovation may be perceived as acceptable but not appropriate and vice versa (Proctor et al., 2011). Fidelity consists of the degree to which an intervention was implemented as it was prescribed in the original protocol (Proctor et al., 2011). More precisely, assessment fidelity – also labelled "procedural fidelity assessment'' (DiGennaro Reed & Codding, 2014) – refers to the respect and adherence to assessment procedures or guidelines (Richardson et al., 2016). In clinical practices, assessment fidelity is important to guarantee the quality of the evaluation carried out and the reliability of the outcomes obtained. One construct for which it is important to measure with fidelity is social vulnerability.

Some adults with ID, whether younger or older, men or women, are more socially vulnerable than others (Nettelbeck & Wilson, 2002). Social vulnerability is a dynamic, fluid, and multidimensional construct (Petitpierre & Tabin, in press). Currently, social vulnerability is defined as the ability to detect or avoid victimization, e.g., physical or sexual assault, financial abuse, or psychological abuse, like bullying or coercion (Fisher et al., 2018). Social vulnerability is not an intrinsic factor of ID. Any individual, including a person with ID, learns to independently navigate potential challenges and thus decrease their social vulnerability through exposure to social situations (Seward et al., 2018). Social vulnerability also depends on the presence or absence of certain risk or protective factors. Studying social vulnerability in adolescents and adults with Williams Syndrome, Down Syndrome, and Autism Spectrum Disorder, Fisher et al. (2018) noted the role of friendships in social vulnerability. Indeed, those with lower education and fewer friends also had less awareness of risk, increasing their social vulnerability. Social vulnerability also appears to tap into some skill over and above either IQ, social intelligence or adaptive behavior (Fisher et al., 2018; Sofronoff et al., 2011; Wilson et al., 1996). Social vulnerability does not systematically lead to abuse, but it indicates a likelihood of being abused. In comparison with other variables such as anxiety, anger, behavior problems and social skills, social vulnerability was found to be the best predictor of bullying victimization for children with Asperger Syndrome (Sofronoff et al., 2011). Likewise, social vulnerability distinguished victims of interpersonal violence (assault, sexual assault, robbery, financial exploitation, break-in) from non-victims in the previous year in a sample of adults with ID (Wilson et al., 1996)

The assessment of social vulnerability aims to evaluate the ability of a person to detect potentially harmful interpersonal situations (Seward et al., 2018). People with disabilities, particularly those with intellectual disabilities, are significantly more at risk of being victims of interpersonal violence than the general population (Dion et al., 2011; Sullivan & Knutson, 2000). Not all social institutions providing services for adults with ID in Switzerland have a psychologist on staff: the assessments are sometimes carried out by educational staff or the team/center manager. In current clinical practices, the social vulnerability assessment of adults with ID is usually conducted by special education practitioners as clinical observations (Petitpierre & Tabin, in press). This evaluation format leaves, however, a lot of room for the subjectivity of the observers. In this context, standardized tools, like self-reported and/or informant-rated tests, are welcome as they offer more uniform evaluation criteria. Together with clinical observations, they provide a more complete and nuanced picture of a person's strengths and limitations. The Social Vulnerability Test-22 items [TV-22] is currently the only social vulnerability measure validated and specifically designed for adults with ID available in French (Tabin et al., 2021). The test is composed of twenty-two illustrated vignettes mimicking real-life social situations. In each item (vignette), the person being assessed is asked to advise a third person (Marie or Pierre, according to the gender they identify with) facing a social risk (theft, verbal or physical aggression, sexual abuse, inappropriate requests or attempts at manipulation).

The TV-22 results from the cross-cultural adaptation of the Test of Interpersonal Competences and Personal Vulnerability [TICPV] from Wilson et al. (1996). In the original study, Wilson et al. (1996) analyzed the reliability and the validity of the TICPV in a sample of 40 adults with ID. Internal consistency (Cronbach's $\alpha = .72$) and test– retest reliability ($r\ (20) = .72, p < .001$) of the TICPV were good; further studies confirmed the good discriminatory validity with larger samples (Murphy & O'Callaghan, 2004; Nettelbeck & Wilson, 1995; Nettelbeck et al., 2000). The TV-22 is the French-language, enhanced (+ 2 items) and accessible version (i.e., following easy-to-read and understand principles, Inclusion Europe, 2009) of the TICPV. The validation was performed on a sample of 29 adults with ID. Internal consistency (Cronbach's $\alpha = .89$; McDonald's $\Omega = .93$) and test-retest reliability ($r_s(29) = .81, p < .01$) of the TV-22 were good. Inter-rater reliability was calculated between two independent coders from the research team with Cohen's κ and ranged from .651 to 1 ($p < .001$) i.e., from substantial to perfect agreement according to Landis and Koch (1977) benchmarks.

In the present research, we aim to evaluate the implementation outcomes (acceptability, appropriateness, assessment fidelity) of the TV-22 as administered by special education practitioners (psychologists, educators, team/center managers). More precisely, we examine the influence of some personal factors in relation to assessment fidelity and acceptability of the test. We further explore the appropriateness, i.e., the perceived relevance and challenges related to the compatibility of the test to the participants' practice. These analyses will highlight facilitators and barriers to implementation outcomes, providing information on the modifications required to ensure the adequate use of the TV-22 and its dissemination.

2. Material and methods

This study is the second strand of a mixed methods test validation project. The whole project lasted from February 2018 to August 2020 (Petitpierre, 2018); data for this second strand were collected in Spring 2019. The results of the first strand, focused on the psychometric properties of the test, have been detailed elsewhere (see (Tabin et al., 2021). The project was approved by the Swiss Ethics Committee on research involving humans (Protocol N°2016-01480) and funded by the Swiss National Science Foundation (Project N° 176196).

2.1. Participants

Thirty-one practitioners were identified by the heads of eight social institutions for adults with ID in the French-speaking part of Switzerland. Participants were primarily female, with more than 15 years of experience in the field. They were divided into two groups (see *Procedure* section for further details). Both groups were invited to a training session on the use of the TV-22. At the end of the training session, each practitioner completed an informed consent document and a short survey. The short survey included questions on demographic information, experience and general attitude towards assessment tools (see Table 1).

2.2. Material

The participants were provided with the Social Vulnerability Test-22 items [TV-22] and its associated materials.

2.2.1. Administration of the test

(1) An instructor's manual briefly introduces the concept of social vulnerability, the ethical precautions to follow when using the test and then describes the instructions for administration of the test as well as the procedures for scoring and interpreting the respondent's responses. The main conditions for administration of the test are the following: each person is assessed individually; there is no time limit per item; the length of time to administer the full set of TV-22 vignettes is flexible. It can vary from one respondent to another. The results of the validation procedure indicate an average duration of 70 min (min. 43 min; max.132 min).

Table 1
Participant demographics, experience and attitude towards assessment tools (n = 31).

Characteristic	Group 1 (regular training)		Group 2 (enhanced training)		Total	
	n	%	*n*	%	*n*	%
Gender						
Male	2	12.5	3	20.0	5	16.0
Female	14	87.5	12	80.0	26	84.0
Age						
20–29	2	12.5	3	20.0	5	16.0
30–39	7	43.7	5	33.3	12	39.0
40–49	2	12.5	4	26.7	6	19.0
50–60	5	31.3	3	20.0	8	26.0
Profession						
Educator	6	37.5	5	33.3	11	35.0
Psychologist	4	25.0	4	26.7	8	26.0
Team/Center Manager	6	37.5	6	40.0	12	39.0
Years of experience in the field of ID						
2–5	3	18.7	5	33.3	8	26.0
6–10	3	18.7	3	20.0	6	19.0
11–15	5	31.3	–	–	–5	16.0
> 15	5	31.3	7	46.7	12	39.0
Experience in using assessment tools with people with ID						
Never used	4	25.0	1	6.7	5	16.0
Rarely used	1	6.3	5	33.3	6	19.5
Sometimes used	9	56.3	4	26.7	13	42.0
Often used	1	6.3	5	33.3	6	19.5
Very often used	1	6.3	–		1	3.0
Point of view on the usefulness of assessment tools						
Not at all important	–	–	–	–	–	–
Low importance	–	–	–	–	–	–
Neutral	3	18.7	6	40.0	9	29.0
Very important	9	56.3	9	60.0	18	58.0
Extremely important	3	18.7	–	–	3	10.0
No answer	1	6.3	–	–	1	3.0

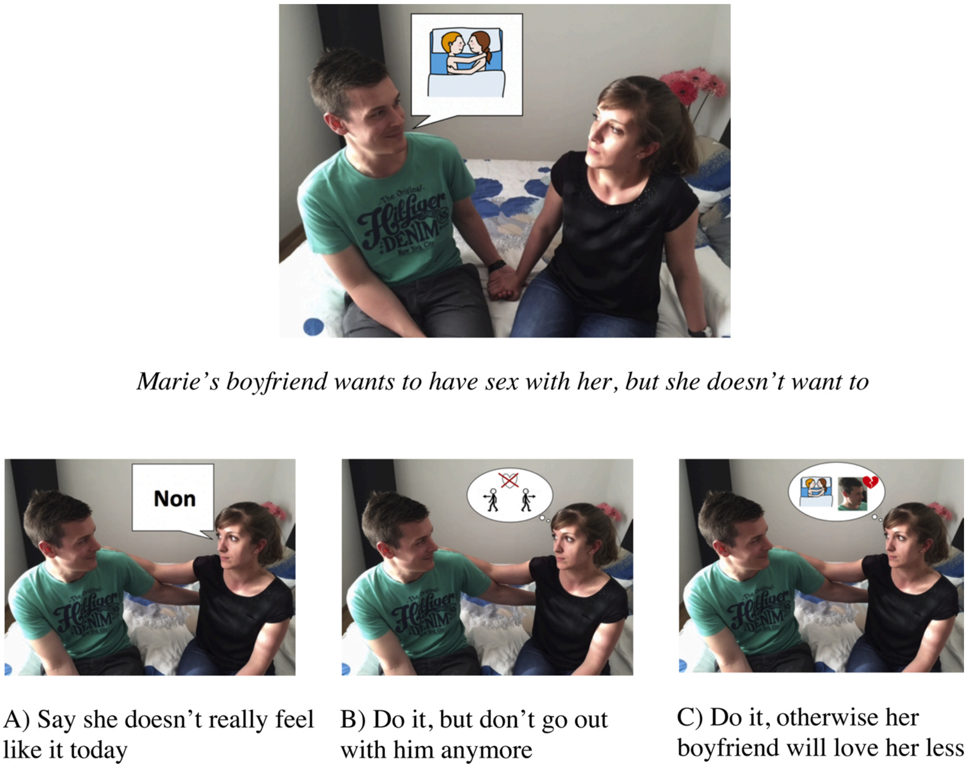
(2) The test is composed of 22 items (88 illustrated vignettes in total). For each vignette introduced, the respondent is first asked to reformulate the situation (What is happening here?). This reformulation requirement aims to check the person's understanding of each vignette. The test, which is administered on a computer (PowerPoint format), consists of two parts following each other. In Part A, the situation (vignette) is presented and the assessor asks for the respondent's spontaneous answer (What do you advise Pierre/Marie to do or say?). In Part B, the vignettes from Part A are followed by three possible answers (a, b, c, also illustrated vignettes). The respondent is then asked to choose the most cautious of the three available answers (see Fig. 1). For each answer given, whether spontaneous answer (part A) or the most cautious answer (part B), the respondent has to justify the proposed strategy (Why should Pierre/Marie do that? (A); Why do you think it is the 'most cautious' answer? (B)).

(3) One reported answers sheet and two scoring sheets (Sheet A; Sheet B) go along with the test and the instructor's manual. When administering the test, the assessor reports all of the respondent's answers on the "Answer Sheet".

2.2.2. *Scoring and interpretation of the test results*

Once the test is performed, the assessor qualitatively analyzes the respondent's spontaneous answers given (Part A) on "Sheet A". The scoring of spontaneous answers from part A provides a first – qualitative – picture of the person's abilities and/or limitations in recognizing the risk present in the situation and their ability to propose strategies to mitigate the risk (for more details about Part A, (see Tabin, 2020) The answers given can be categorized in two major categories: either the respondent identifies and handles the risk (1) or s/he does not (2). When the respondent answers that there is a problem in the situation presented (risk awareness), and mentions one or more protection strategies, their answer is coded according to five sub-categories ((1)A self-protection, (1)B indirect protection; (2)C no strategy; (2)D emotional strategy; (2)E avoidance strategy). Cohen's κ was run to determine agreement between two independent coders from the research team on the categorization of spontaneous responses collected in Part A in 20 % of the data ($n = 6$). Cohen's κ ranged from .700 to 1, i.e., from substantial to perfect agreement according to Landis and Koch (1977) benchmarks.

The results from Part B provide a score of social vulnerability and are reported on 'Sheet B'. The 'most cautious' answer is scored 1, whereas the other options are scored 0. The higher the score (22 points maximum), the more capable the person is of choosing strategies that tend to protect Pierre/Marie. Inter-rater agreement between the two independent coders from the research team on the categorization of multiple-choice responses collected in Part B (on 20 % of the data) was similar to Part A and ranged from κ = .667 to 1.



Marie's boyfriend wants to have sex with her, but she doesn't want to

A) Say she doesn't really feel like it today

B) Do it, but don't go out with him anymore

C) Do it, otherwise her boyfriend will love her less

Fig. 1. Example of a TV-22 item before the implementation study (female version, item N°17,English translation).
Note. The assessor asks: what is happening here? What do you advise Marie to do or say? Why? The assessor then shows each of the three following options (A, B, C) and asks: what is the 'most cautious' answer? Why?

2.3. Procedure

All participants were invited to a training session of approximately 90 min on the use of the TV-22. They were divided into two groups according to a stratified random distribution principle by function. The purpose of dividing the participants into two groups was to assess which kind of training was most suitable for an optimal appropriation of the principles of administration of the TV-22.

The first group ("regular training", $n = 16$) received a general introduction on the origin and goals of the test. Practitioners were given the instructor's manual and test materials (a USB flash drive with the test, reported answers sheet, scoring Sheet A, scoring Sheet B).

The second group ("enhanced training", $n = 15$) followed the same procedure, with the difference that they received a tutorial in addition to the instructor's manual. This tutorial describes the administration of the test in video format: it consists of a PowerPoint file with recorded comments (21 min) describing the instructions for administration of the test, and the procedures for scoring and interpreting the respondent's responses. It also includes a short video (4 min) of an administration example with an assessor and a respondent using the TV-22.

At the end of the training session, each practitioner was required to administer the TV-22 with an adult with a mild to moderate ID with whom they were working, to record the interview on audio tape, score the responses of the person they had assessed, and then send the recording of the test and the scoring results to the research team within one month. Two out of 31 practitioners (one from each group) encountered technical difficulties and were unable to record their interview. Twenty-nine interviews were fully transcribed by the research team.

Semi-structured interviews were conducted with practitioners to elucidate their experience using the TV-22, with a focus on identifying barriers and facilitators to implementation. Interview duration was approximately 40 min. One member of the research team conducted the interview, whereas the other reported what the participant said. The interviews were also recorded, to be able to go back to the exact words used if needed. The interview was divided into three parts:

- The first part addressed the assessment fidelity of the TV-22 and, more precisely, correct administration and use of the materials provided: 12 questions were dichotomous (yes/no type) (e.g., assessor and respondent were sitting next to each other; they were alone in a room; the reported answers sheet was used during the administration).
- The second part addressed the acceptability of the TV-22: five open-ended questions were asked (e.g., *What did you find easy or, conversely, difficult during the administration of the TV-22? What do you think about the 22 vignettes of the test?*) and three 5-point Likert-type scale questions ranging from very easy (1) to very difficult (5) about the complexity of administering the TV-22, the complexity of the scoring; and the complexity of the interpretation of the social vulnerability scores.
- Finally, five open-ended questions were about the appropriateness of the TV-22 (e.g., *Why do you think that the TV-22 will (not) be useful for your clinical practice?*).

2.4. Analysis

In order to evaluate the assessment fidelity of the TV-22, quantitative analyses (using SPSS, version 26) were conducted. The evaluation of the fidelity was based on (1) analysis of the interview transcripts between the practitioner and the adult with ID, the scoring Sheet A and the scoring Sheet B completed by the practitioner; and (2) the answers given by the practitioner to the first 12 questions of the semi-structured interview.

Three dimensions of the assessment fidelity were analyzed: compliance with the administration conditions (12 criteria), conformity

Table 2
Structural-procedural critical components evaluated for assessment fidelity.

Dimension	Criteria		Rating
Compliance with the administration conditions	12 criteria	5 criteria: Compliance with the general administration conditions (e.g., assessor and respondent were alone in a room); 7 criteria: The use of prescribed material (e.g., the reported answers sheet was used during the administration);	Dichotomous (yes, criteria respected = 1/no = 0)
Conformity to the administration procedure	4 criteria * 22 items	1*22: Reformulation asked (*What is happening here?*)	Trichotomous (yes = 2/partially = 1/no = 0)[a]
		1*22: Part A question asked (*What do you advise Pierre/ Marie to do or say?*)	Dichotomous
		1*22: Justification asked (*Why should Pierre/Marie do that?*)	Dichotomous
		1*22: Part B question asked (*What is the 'most cautious' answer* (A, B or C)?	Dichotomous
Scoring the items	2 criteria * 22 items	1*22: Open-ended questions (Part A) 1*22: Multiple-choice questions (Part B)	Agreement between the scoring given by the practitioners and the member of the research team

[a] *Note*. The reformulation was sometimes skipped by the practitioners when the respondents spontaneously reacted (e.g., laughed, said ''oh my god''). Because it reflects to some extent an understanding – which is the aim of the reformulation requirement –, if the practitioner did not explicitly ask for the reformulation, those reactions were coded as 'partially' fulfilled (1 point).

Table 3
Inter-rater reliability scores of open-ended questions.

Items	Group 1 (regular training)			Group 2 (enhanced training)			Total		
	n	κ	*p*	*n*	κ	*p*	*n*	κ	*p*
1	12	.652	<.001	13	−.054	.657	25	.442	<.001
2	10	.714	<.001	13	.490	<.001	23	.568	.001
3	11	.421	.002	13	.725	<.001	24	.596	<.001
4	12	.526	.002	13	.755	<.001	25	.660	<.001
5	12	.339	.005	13	.253	.060	25	.316	.001
6	10	.375	.062	12	.167	.157	22	.361	.003
7	12	.442	.029	11	−.065	.621	23	.228	.039
8	12	.435	.002	12	.721	<.001	24	.544	<.001
9	11	.457	.014	12	.200	.190	23	.417	<.001
10	11	.353	.018	13	.278	.038	24	.314	.002
11	10	.778	.003	13	.217	.222	23	.448	.003
12	12	.314	<.001	13	1	<.001	25	.497	<.001
13	11	.233	<.001	13	.711	<.001	24	.458	<.001
14	12	.377	.015	13	.325	.029	25	.353	.001
15	12	.400	.000	12	.700	<.001	24	.557	<.001
16	10	.022	.747	13	.443	.001	23	.240	.001
17	12	.642	.001	13	.536	.001	25	.593	<.001
18	11	.667	.001	12	.294	.020	23	.576	<.001
19	10	.600	<.001	13	.669	<.001	23	.641	<.001
20	10	.302	.074	13	.639	.001	23	.449	<.001
21	11	.522	.017	13	.494	.001	24	.506	<.001
22	11	1	.001	13	.297	.015	24	.217	.071

to the assessment procedure (4 criteria*22 items), and scoring of the items (2 criteria*22 items), comprising a total of 144 structural-procedural critical components (see Table 2 for details). These components were coded independently by two members of the research team for each fully completed test administration ($n = 29$). Inter-rater agreement, calculated on 20 % of the data ($n = 6$) for the 144 components, ranged from κ = .667. to 1, i.e. from substantial to perfect agreement according to Landis and Koch (1977) benchmarks.

In order to assess the factors influencing the assessment fidelity and the acceptability of the test (e.g., gender, age, profession), quantitative analyses (using SPSS, version 26) were performed. Qualitative analyses (using NVivo, version 12) were conducted to assess the appropriateness of the test. We conducted a thematic analysis (Paillé & Mucchielli, 2012). This analysis requires that segments of text ranging from a phrase to several paragraphs were assigned codes based on a priori (i.e., from the semi-structured interview guide) themes. Afterwards, these themes were further analyzed to check whether they repeat from one segment to another and how they intersect, join, contradict or complement each other (Paillé & Mucchielli, 2012). This analysis constructs a

Table 4
Inter-rater reliability scores of multiple-choice questions.

Items	Group 1 (regular training)			Group 2 (enhanced training)			Total		
	n	κ	*p*	*n*	κ	*p*	*n*	κ	*p*
1	14	.827	<.001	14	1	<.001	28	.874	<.001
2	13	.639	.001	14	.641	.001	27	.638	.003
3	15	.464	.002	14	.641	.001	29	.540	<.001
4	14	.877	<.001	14	.481	<.001	28	.816	<.001
5	15	.694	<.001	14	1	<.001	29	.843	<.001
6	15	.697	<.001	14	1	<.001	29	.773	<.001
7	14	1	<.001	14	.825	<.001	28	.894	<.001
8	14	1	<.001	14	.641	<.001	28	.844	<.001
9	14	1	<.001	14	1	<.001	28	1	<.001
10	14	.720	<.001	14	1	<.001	28	.851	<.001
11	15	.571	.003	13	.863	<.001	28	.723	<.001
12	15	.868	<.001	14	1	<.001	29	.920	<.001
13	15	1	<.001	14	1	<.001	29	1.00	<.001
14	15	.727	.002	14	.641	.003	29	.765	<.001
15	15	.737	<.001	13	.527	.020	28	.668	<.001
16	14	.580	<.001	14	1	<.001	28	.704	<.001
17	14	1	<.001	14	1	<.001	28	1	<.001
18	14	1	<.001	14	.770	<.001	28	.873	<.001
19	14	.592	.007	14	.867	<.001	28	.736	<.001
20	15	1	<.001	14	1	<.001	29	1	<.001
21	15	1	<.001	14	1	<.001	29	1	<.001
22	15	1	<.001	14	1	<.001	29	1	<.001

panorama which identifies the major trends of the phenomenon, here the facilitators and barriers to implementation outcomes.

3. Results

Among special education practitioners, psychologists emerged as using the test more rigorously. Assessment fidelity was related to years of experience in special education, current profession and experience with assessment tools; but not to gender, age or type of training. Acceptability of the measure was not influenced by any of the practitioners' characteristics. The results of the qualitative analyses of the appropriateness revealed the relevance and usefulness of the test as well as the most common challenges faced by the practitioners when using the TV-22: fatigue linked to the respondent being required to reformulate; reporting the respondent's responses during administration of the test; analysis of open-ended questions. Each outcome provides valuable information on the modifications needed to ensure the adequate use of the TV-22 and promote its dissemination.

3.1. Assessment fidelity

Assessment fidelity has been analyzed separately for: (1) compliance with the administration conditions (12 criteria = 12 points), (2) conformity to the administration procedure (4 criteria*22 items, coded either in a dichotomous or trichotomous way = 110 points) and (3) scoring of the items (2 criteria*22 items = 44 points).

(1) Administration conditions

The results show that the conditions for administration were followed by most participants. The 12 criteria assessed (e.g., quiet room; Sheet B was used as described in the user's manual) were respected to a great extent (M = 96.5 %, with a range between 87.1%–100%).

(2) Conformity to the administration procedure

Regarding the administration of the items (i.e., asking the questions to assess the respondent's social vulnerability), practitioners obtained, on average, a score of 69 (SD = 26.09) – a score which displays a relatively low adherence to the administration guidelines. A maximum score of 110 per administration can be obtained; the higher the score, the better the administration, as it reflects that the practitioner asked each of the questions required to be asked in order to assess the respondent's social vulnerability, on each of the 22 items of the test. The maximum total fidelity score achieved is 109. Seven practitioners obtained fewer than 50 points; 5 (17 % of the sample) obtained 44 points or less, which means that they asked fewer than 50 % of the questions required to perform the assessment.

(3) Scoring of the items

The inter-rater reliability (i.e., the rating given by the research team member versus the rating given by the practitioner) was calculated separately for open-ended questions (Part A) and multiple-choice questions (Part B).

For open-ended questions the inter-rater reliability, calculated with Cohen's κ, is fair for 7 out of 22 items ($\kappa < .400$); moderate for 13 items; and substantial for 2 items ($\kappa > .610$) (Landis & Koch, 1977; see Table 3 for detailed Cohen's κ results). Cohen's κ is globally unsatisfactory for the open-ended questions, as it ranges below what could be expected at an individual level and is also way below the agreement found between two independent coders from the research team, which ranged between .700 and 1 (see section *Scoring and interpretation of the test results*).

We can also note that with regard to compliance with the guidelines underlying Part A of the test (open-ended questions), the number of participants varies and never reaches 29 (i.e., the total number of participants expected for these analyses): some open-ended questions were systematically and fully skipped by some practitioners and were thus coded as 'missing data' and excluded from reliability analysis on the scoring fidelity.

As regards the multiple-choice questions, except for item 3 ($\kappa = .540$; $p < .001$), the inter-rater reliability ranged between .638 and 1, i.e., substantial to excellent agreement (Landis & Koch, 1977; see Table 4 for detailed Cohen's κ results). This result is also similar to the agreement found between two independent coders from the research team, which ranged between .667 and 1 (see *Scoring and interpretation of the test results section*)

3.2. Factors influencing implementation outcomes

The assessment fidelity, more specifically the administration of the items (i.e., asking all the questions required to be asked in each item in order to assess the respondent's social vulnerability – What is happening here? What should Pierre/Marie do? Why? Which is the most cautious answer?) was influenced by the profession, the years of experience in the field of intellectual disabilities and the experience with assessment tools with people with ID.

Psychologists administered the test more rigorously than educators, who, in turn, complied more closely with the criteria than team or center managers ($n = 29$, $H(2) = 9.324$, $p = .009$, $\eta^2 = 0.282$, $d_{cohen} = 1.252$). Participants who reported using – or having used – tools to conduct assessments with people with ID are more rigorous in their administration ($n = 29$, $H(4) = 9.917$, $p = .042$, $\eta^2 = 0.247$, $d_{cohen} = 1.144$). On the other hand, years of experience with ID are negatively correlated with assessment fidelity ($n = 29$, $r_s = -.481$, p

$< .01$). This means that the more experience participants have with people with ID, the less formal they were and the less they complied with the recommended guidelines.

With respect to the acceptability of the materials, with an overall average of 11.12 out of 15 ($SD = 2.09$), participants found the TV-22 user-friendly and easy to administer, score and interpret. The overall perception of the acceptability of the test was not influenced by the personal characteristics of the participants (age, gender, profession, years of experience in the field of ID), nor by their experience with assessment tools. The instruction conditions (group 1 –regular training– versus group 2 –enhanced training), influence the perceived appropriateness level: participants who benefited from the tutorial and the manual tend to find the tool more user-friendly than those who benefited only from the manual ($U = 81$, $z = -1.585$, $p = .065$ (one-tailed), $r = .28$, $\eta^2 = 0.077$, $d_{cohen} = 0.576$).

3.3. *Appropriateness*

3.3.1. *Relevance and compatibility of the test to the practice*

Twenty-six participants highlighted the relevance of the content of the TV-22 items, like this practitioner who reports that "the test refers to many situations that have been talked about or experienced within the institution, so they are truly vignettes that correspond to reality, to observations". All participants acknowledged the compatibility of the test for their practice, albeit for different reasons: (1) as an assessment tool to improve support of adults with ID (17 occurrences); (2) as a support to stimulate discussions on certain themes/risks (10 occurrences); as a tool for assessing progress (test-retest) (5 occurrences); and finally, (3) as a "mediator" between teams and/or with legal representatives when there are different opinions on a person's ability to manage socially risky situations (5 occurrences), like this practitioner who said: "it's an interesting tool for exchanges between teams: sometimes there are disagreements [around the question of risks versus self-determination], a tool like this makes things clearer and more objective".

3.3.2. *Challenges faced when using the test*

During the administration of the test, participants reported having faced difficulties relating to assessor neutrality, like this participant who reports "it's hard not to comment when the respondent chooses a non-cautious answer". Participants have also reported some difficulty in abiding by the test administration instructions. More specifically, some faced the dilemma of choosing a literal – but potentially robotic or unnatural – way of administering the test and a more spontaneous – but unstandardized – one. A majority of participants found the test (too) long to administer or reported that the respondent seemed to be annoyed by the fact that s/he had to reformulate what was going on at each new vignette (i.e., the "what is happening here" question). Nineteen participants also stated that they struggled with reporting the respondent's answer, mostly either because there was not enough space on the reported answers sheet or because it interrupted the thread of the interview, reducing the opportunity to deepen the discussion with the respondent.

The most common challenge faced by participants when using the TV-22 was the analysis of answers to open-ended questions (Part A) – which confirms the results from reliability analysis described above. Some participants also faced difficulties in interpreting the respondents' scores. Table 5 summarizes the challenges reported by participants when using the TV-22.

4. Discussion

4.1. *Challenges for implementation: modifications to the materials and instructions*

The qualitative and quantitative results highlight several potential challenges for implementation of the TV-22. With an intentional focus on retaining the core content of the test, a series of modifications and improvements were made to the TV-22 materials and instructions to address these challenges and reduce barriers to dissemination (see Table 6).

First of all, it appears to be necessary to offer half-day training sessions in TV-22 administration for practitioners who are not psychologists. This adjustment aims to improve fidelity. Furthermore, to ensure assessment fidelity, every question that the assessor has to ask is now written directly on the test so that it can simply be read (see Fig. 2).

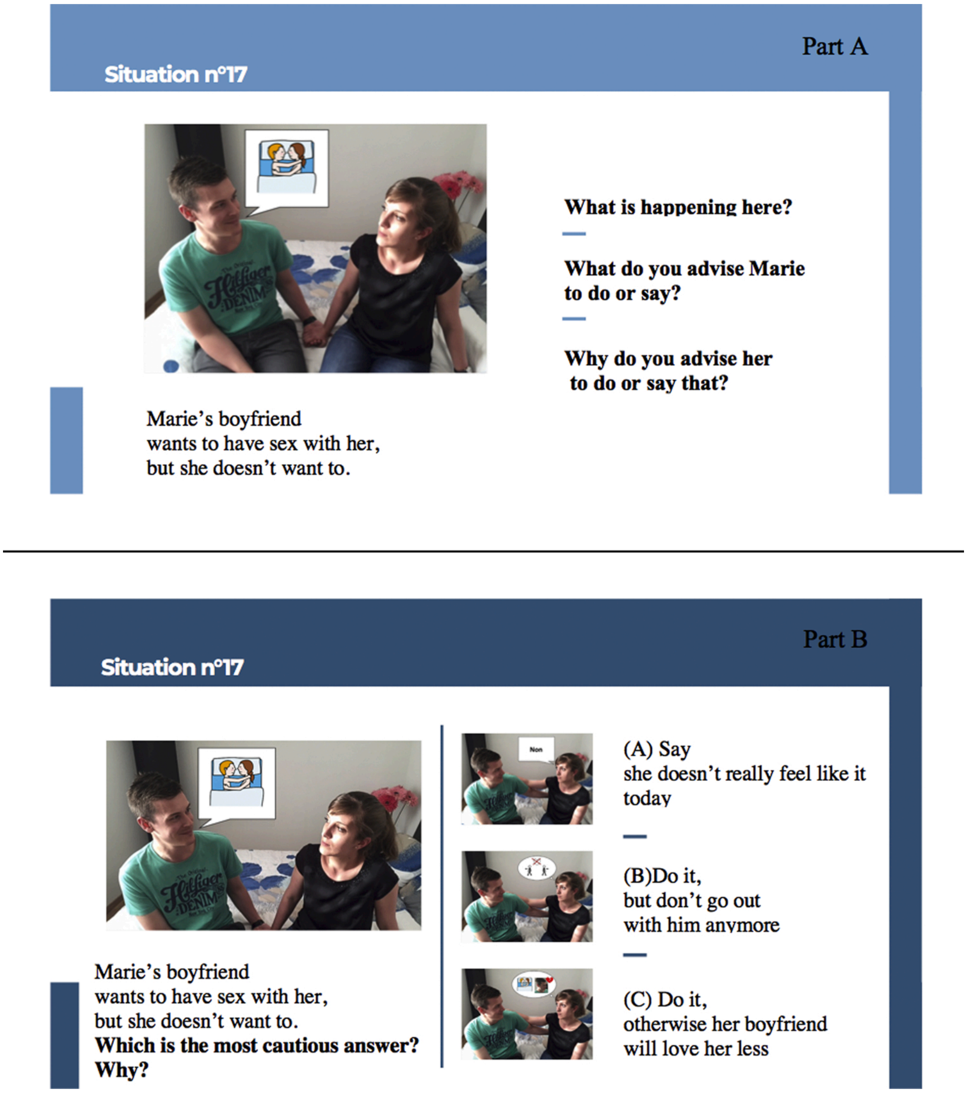
In order to enhance acceptability, two short demonstration videos (Part A and Part B) have been made available online.

Table 5
Challenges reported by practitioners when using the TV-22 (n=31).

Test phase	Challenge reported	Frequency, *n* (%)		
		Group 1 (regular training)	Group 2 (enhanced training)	Total
Administration	Assessor Neutrality	2	5	7 (23)
	Robotic *versus* unstandardized administration dilemma	2	2	4 (13)
	Fatigue with requirement to reformulate to check if the person has understood the situation	7	7	14 (45)
	Length of the test administration	6	5	11 (35)
	Report of Responses	9	10	19 (63)
Scoring	Analysis of Part A responses	11	10	21 (68)
Interpretation	Interpretation of Part B scores	8	5	13 (42)

Table 6
Summary of key modifications to the TV-22.

Implementation outcome	Challenge identified	Type of modification
Fidelity	• Assessment fidelity not always guaranteed; • Respect of Part A core component (open-ended questions; coding of responses) not guaranteed	• Training recommended for non-psychologists; • Questions written directly on the test itself;
Acceptability	• Tutorial could enhance acceptability	• Demonstration videos made available online
Appropriateness	• Assessor Neutrality; Robotic *versus* unstandardized administration dilemma; • Report of Responses; analysis of Part A responses; interpretation of Part B scores; • Fatigue with requirement to reformulate to check if the person has understood the situation; • Length of the administration	• Additional information added in the instructor's manual (administration, scoring, results interpretation); • Record audio of test administration to check response reporting and scoring; • Reformulation required for each new item (i.e., 22 times) – and not each new vignette (i.e., 88 times); • Part A and Part B performed at different times



Situation n°17

Part A

What is happening here?

What do you advise Marie to do or say?

Why do you advise her to do or say that?

Marie's boyfriend wants to have sex with her, but she doesn't want to.

Situation n°17

Part B

Marie's boyfriend wants to have sex with her, but she doesn't want to.
Which is the most cautious answer? Why?

(A) Say she doesn't really feel like it today

(B)Do it, but don't go out with him anymore

(C) Do it, otherwise her boyfriend will love her less

Fig. 2. Example of a TV-22 item after the implementation study (female version, item N°17, English translation).
Note. The assessor or the respondent successively reads the scenario, the question and the optional answers. Then the respondent answers the questions and justifies their answer. Approximately one week later, both meet again to take part B of the test.

The poor inter-rater reliability result for the coding of part A (open-ended question) and the qualitative results from the appropriateness analysis have led us to refine considerably the scoring instructions for open-ended questions in the instructor's manual. It is also now recommended to record the audio of the test administration to be able to check response reporting and scoring afterwards.

The results of the appropriateness analyses also highlighted the need to add important details in the manual. Two changes were also made in the instructions to counter the challenges linked to the reformulation and the length of the administration (see Table 6 for details).

4.2. *Assessment fidelity: differences between practices and assumption of an error-free use*

Reliable assessments are important to guarantee quality of care. Yet, although reliable assessments are developed and validated through research processes, some gaps between research and practice remain. Newly developed assessments may sometimes get lost in the so-called "leaky" research pipeline (De Geest et al., 2020). By evaluating implementation outcomes related to the use of the TV-22, a newly validated test, we aimed to identify the potential challenges related to its use, and thus overcome translation barriers between research and practice. The results of this study indicate a good appropriateness and a reasonable acceptability, – but nevertheless a low assessment fidelity of the TV-22 by some practitioners.

The picture that emerges from the fidelity analysis is one of different assessment practices depending on one's profession. We have noticed a difference between current profession and adherence to assessment guidelines. On the one hand, it is not surprising that psychologists, who are trained and used to performing assessments with standardized procedures, are more rigorous in their administration of the test than educators and head/center managers. On the other hand, it is striking that 25 % ($n = 7$) of the practitioners complied with less than one half of the assessment guidelines – which definitely questions the quality of the social vulnerability assessment performed. Similarly, in a study examining administration and interpretation of reading tests, Reed and Sturges (2013) report 8% of extreme lack of fidelity in test administration and 91 % of accurate test administration – but still with correctable errors. Nevertheless, research on assessment fidelity remains scarce (Reed et al., 2014; Richardson et al., 2016). In a literature review on assessment fidelity in special education, more precisely in reading interventions among children, Reed et al. (2014, p. 310) note that "there appears to be an assumption that the measures are always used in an error-free fashion'' (p. 310). This assumption leads to low reports of assessment fidelity in implementation researches, despite its important role for interpretation of the results.

4.3. *Implementation of an innovation and the need to make adaptations: tensions and perspectives*

The results relating to the assessment fidelity, appropriateness and acceptability of the TV-22 by practitioners have led to several modifications of its associated materials. While the preliminary validation study of the test reported good psychometrics properties (Tabin et al., 2021), one may wonder if the modifications carried out had an impact on its reliability. This tension between the implementation of an innovation and the need to make (local) adaptations is not new (DeGue et al., 2020; Elliott & Mihalic, 2004). It is part of a debate between a balance-adaptation mindset and those who consider that any bargaining away of fidelity will negatively impact the effectiveness of the program or the innovation (Elliott & Mihalic, 2004). In this study, the modifications were not made to the content of the test itself but to its associated materials; they aimed to improve clarity of the guidelines and overcome assessment fidelity issues observed. These modifications should thus increase assessment fidelity, supporting its reliability rather than being a threat to it. Likewise, the modifications made should enhance appropriateness and acceptability of the test, supporting its dissemination. Nevertheless, additional studies are required to assess the impact of these modifications and explore whether aspects of training and administration can be manipulated to improve implementation outcomes, and more specifically assessment fidelity – an area of research that has, at present, received little attention (Richardson et al., 2016).

4.4. *Limitations*

The present research has several limitations worth noting. The small size and heterogeneity of the sample (psychologists, educators, and team/center managers) may have hindered finding group differences. A study with a larger sample size would have been necessary to explore this type of effect. Furthermore, we did not collect data on the characteristics of the person being evaluated (e.g., severity of ID, age, comorbidity). We cannot rule out that these characteristics have had an impact on the measured outcomes, thus raising the question of the generalization of the results. Future work should look to further explore the relationships between the characteristics of the person being evaluated and the TV-22 implementation outcomes in a larger sample of practitioners. Additionally, results focusing on acceptability and appropriateness come from semi-structured interviews that we conducted with the practitioners. As the same research team validated the test and performed the acceptability and appropriateness evaluations, social desirability and politeness bias may have prevented practitioners from raising criticism of the test. To prevent this bias, future research should use validated measures to assess implementation outcomes – nevertheless, despite recent improvement (Mettert et al., 2020), implementation outcomes instrumentation is underdeveloped (Lewis et al., 2015), the first steps should consist of developing and validating psychometrically sound implementation outcome instruments in French.

5. Conclusions

Potential adopters should have the opportunity to try out the innovation, while having sufficient autonomy to "work around" and refine the innovation to improve its compatibility with the intended purpose (Greenhalgh et al., 2004). Thanks to the collaboration of practitioners, modifications were made to the newly developed test, in order to increase its adequacy to the practice's needs. This study delivers insight into the implementation processes of a new measure in the special education field. The results highlight the importance of evaluating implementation outcomes when a new measure is developed to ensure its appropriateness and correct use by stakeholders – and prevent it from getting lost in the "leaky" research pipeline.

Funding

This study was supported by the Swiss National Science Foundation [grant number 176196].

CRediT authorship contribution statement

Mireille Tabin: Project administration, Investigation, Data curation, Formal analysis, Writing - original draft. **Cindy Diacquenod:** Project administration, Investigation, Data curation, Writing - review & editing. **Geneviève Petitpierre:** Conceptualization, Methodology, Supervision, Funding acquisition, Writing - review & editing.

Declaration of Competing Interest

The authors report no declarations of interest.

References

De Geest, S., Ziga, F., Brunkert, T., Deschodt, M., Zullig, L. L., Wyss, K., & Utzinger, J. (2020). Powering Swiss health care for the future: Implementation science to bridge "the valley of death". *Swiss Medical Weekly*. https://doi.org/10.4414/smw.2020.20323

DeGue, S., Le, V. D., & Roby, S. J. (2020). The Dating Matters ® Toolkit: Approaches to increase adoption, implementation, and maintenance of a comprehensive violence prevention model. *Implementation Research and Practice, 1*, 1–12. https://doi.org/10.1177/2633489520974981

DiGennaro Reed, F. D., & Codding, R. S. (2014). Advancements in procedural fidelity assessment and intervention: Introduction to the special issue. *Journal of Behavioral Education, 23*(1), 1–18. https://doi.org/10.1007/s10864-013-9191-3

Dion, J., Matte-Gagné, C., Tourigny, M., & Gaudreault, L. (2011). Les enfants avec retard sont plus exposés à la maltraitance et relèvent davantage des services de la protection de la jeunesse [Children with developmental delays are at greater risk of maltreatment and are more likely to be in the care of child welfare services]. *Enfance, 4*, 421–443.

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology, 41*(3–4), 327–350. https://doi.org/10.1007/s10464-008-9165-0

Elliott, D. S., & Mihalic, S. (2004). Issues in disseminating and replicating effective prevention programs. *Journal of Prevention Science, 5*(1), 47–53. https://doi.org/10.1023/b:prev.0000013981.28071.52

Fisher, M. H., Shivers, C. M., & Josol, C. K. (2018). Psychometric properties and utility of the social vulnerability questionnaire for individuals with intellectual and developmental disabilities. *Journal of Autism and Developmental Disorders*. https://doi.org/10.1007/s10803-018-3636-4

Greenhalgh, T., Robert, G., Macfarlane, F., Bate, P., & Kyriakidou, O. (2004). Diffusion of innovations in service organizations: Systematic review and recommendations. *The Milbank Quarterly, 82*(4), 581–629. https://doi.org/10.1111/j.0887-378X.2004.00325.x

Inclusion Europe. (2009). *L'information pour tous, règles européennes pour une information facile à lire et à comprendre [Information for all, European standards for making information easy to read and understand]*. https://easy-to-read.eu/fr/european-standards/.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310

Lewis, C. C., Fischer, S., Weiner, B. J., Stanick, C., Kim, M., & Martinez, R. G. (2015). Outcomes for implementation science: An enhanced systematic review of instruments using evidence-based rating criteria. *Implementation Science, 10*(1), 155. https://doi.org/10.1186/s13012-015-0342-x

Mettert, K., Lewis, C., Dorsey, C., Halko, H., & Weiner, B. (2020). Measuring implementation outcomes: An updated systematic review of measures' psychometric properties. *Implementation Research and Practice*. https://doi.org/10.1177/2633489520936644

Murphy, G. H., & O'Callaghan, A. C. (2004). Capacity of adults with intellectual disabilities to consent to sexual relationships. *Psychological Medicine, 34*, 1347–1357. https://doi.org/10.1017/S0033291704001941

Nettelbeck, T., & Wilson, C. (1995). *Criminal victimisation: The influence of interpersonal competence on personal vulnerability*. https://www.criminologyresearchcouncil.gov.au/reports/16-94-5.pdf.

Nettelbeck, T., & Wilson, C. (2002). Personal vulnerability to victimization of people with mental retardation. *Trauma, Violence, & Abuse, 3*, 289–306. https://doi.org/10.1177/1524838002237331

Nettelbeck, T., Wilson, C., Potter, R., & Perry, C. (2000). The influence of interpersonal competence on personal vulnerability of persons with mental retardation. *Journal of Interpersonal Violence, 15*, 46–62. https://doi.org/10.1177/088626000015001004

Paillé, P., & Mucchielli, A. (2012). *L'analyse qualitative en sciences humaines et sociales [Qualitative analysis in the humanities and social sciences]* (2^{e} éd.). Armand Colin.

Petitpierre, G. (2018). Validation du Test de Compétences Interpersonnelles et de Vulnérabilité Personnelle, version Française, Enrichie, Accessibilisée et Informatisée [TCIVP-FEAI]. Projet n°176196 financé par le Fonds National Suisse de la Recherche Scientifique [Validation of the Test of Interpersonal Skills and Personal Vulnerability, French version. Project n°176196 financed by the Swiss National Science Foundation]. http://p3.snf.ch/Project-176196.

Petitpierre, G., & Tabin, M. (2021). From social vulnerability assessment to active prevention measures: a decision-making perspective. In I. Khemka, & L. Hickson (Eds.), *Decision Making by Individuals with Intellectual and Developmental Disabilities: Research and Practice*. Springer. In Press.

Proctor, E., Silmere, H., Raghavan, R., Hovmand, P., Aarons, G., Bunger, A., … Hensley, M. (2011). Outcomes for implementation research: Conceptual distinctions, measurement challenges, and research agenda. *Administration and Policy in Mental Health and Mental Health Services Research, 38*(2), 65–76. https://doi.org/10.1007/s10488-010-0319-7

Reed, D. K., & Sturges, K. M. (2013). An examination of assessment fidelity in the administration and interpretation of reading tests. *Remedial and Special Education, 34*(5), 259–268. https://doi.org/10.1177/0741932512464580

Reed, D. K., Cummings, K. D., Schaper, A., & Biancarosa, G. (2014). Assessment fidelity in reading intervention research: A synthesis of the literature. *Review of Educational Research, 84*(2), 275–321. https://doi.org/10.3102/0034654314522131

Richardson, J. D., Hudspeth Dalton, S. G., Shafer, J., & Patterson, J. (2016). Assessment fidelity in aphasia research. *American Journal of Speech-Language Pathology, 25*(4S). https://doi.org/10.1044/2016_AJSLP-15-0146

Seward, R. J., Bayliss, D. M., & Ohan, J. L. (2018). The Children's Social Vulnerability Questionnaire (CSVQ): Validation, relationship with psychosocial functioning, and age-related differences. *International Journal of Clinical and Health Psychology, 18*(2), 179–188. https://doi.org/10.1016/j.ijchp.2018.02.001

Sofronoff, K., Dark, E., & Stone, V. (2011). Social vulnerability and bullying in children with Asperger syndrome. *Autism, 15*(3), 355–372. https://doi.org/10.1177/1362361310365070

Sullivan, P. M., & Knutson, J. F. (2000). Maltreatment and disabilities: A population-based epidemiological study. *Child Abuse & Neglect, 24*(10), 1257–1273. https://doi.org/10.1016/S0145-2134(00)00190-3

Tabin, M. (2020). Ressources et vulnérabilités des adultes présentant une déficience intellectuelle face aux risques numériques. [Resources and vulnerabilities of adults with intellectual disabilities faced with digital risks]. *Revue francophone de la déficience intellectuelle [The francophone journal of intellectual disability], 30*, 13–24. https://doi.org/10.7202/1075352ar

Tabin, M., Diacquenod, C., De Palma, N., Gerber, F., Straccia, C., Wilson, C., … Petitpierre, G. (2021). Cross-cultural preliminary validation of a measure of social vulnerability in people with intellectual disabilities. *Journal of Intellectual & Developmental Disability*, 1–13. https://doi.org/10.3109/13668250.2020.1793450

Wilson, C., Seaman, L., & Nettelbeck, T. (1996). Vulnerability to criminal exploitation: Influence of interpersonal compe-tence differences among people with mental retardation. *Journal of Intellectual Disability Research, 40*, 8–16. https://doi.org/10.1111/j.1365-2788.1996.tb00597.x