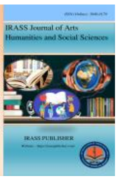
IRASS Journal of Arts, Humanities and Social Sciences
https://irasspublisher.com/journal-details/IJAHSS
ISSN *(Online) 3049-0170*



Achenbach System of Empirically Based Assessment: A multiple applications Tool in a Social Emotional Learning Intervention

Leonidou Pagona[1*], Kartasidou Lefkothea[2]

[*1] PhD student, Department of Educational Policy, University of Macedonia, Greece

[2] Professor, Department of Educational Policy, University of Macedonia, Greece

Corresponding Author Leonidou Pagona

PhD student, Department of Educational Policy, University of Macedonia, Greece

Article History

Received: 11 / 10 / 2025

Accepted: 26 / 11 / 2025

Published: 05 / 12 /2025

Abstract: Objective: A gap seems to exist in research concerning the implementation of Social Emotional Learning interventions for students with disabilities in special education schools. The aim of the particular paper, within the framework of a broader research, is to highlight the multiple applications of Achenbach System of Empirically Based Assessment (ASEBA) in a SEL Intervention in a Special Education School in Greece. Method: A mixed-method research design was applied, as with the combination of quantitative and qualitative methods the weak points of each method are compensated separately. In this paper, though, only the assessment process through ASEBA is going to be presented, so as to point out its multiple applications. In this multiple case study 3 teachers and a parent of each of the 3 students participated. Findings: So, an overall improvement was perceived concerning the behavior of all students, although different perceptions were noticed e.g. in terms of context, as it became evident in the social validity measures. Implication for Practice and Research: ASEBA is a valuable tool for screening procedure, assessing behavior change and social validity in SEL interventions through its multi-informant approach, detailed subscales and normative data. our findings contribute to both in-service and novice teachers' or researchers' knowledge about how to use a multipurpose tool like Achenbach System of Empirically Based Assessment (ASEBA) in the assessment process of a SEL intervention program. Furthermore, an important aspect of our article is that social validity is addressed, which is quite sidelined in research and in practice, even though it constitutes a crucial aspect of any intervention program.

Keywords: *Achenbach System of Empirically Based Assessment (ASEBA); Social Emotional Learning (SEL) intervention program; behavior problems; measurement of change; social validity; intellectual disability.*

How to Cite in APA format: Leonidou, P. & Kartasidou, L. (2025). Achenbach System of Empirically Based Assessment: A multiple applications Tool in a Social Emotional Learning Intervention. IRASS Journal of Arts, Humanities and Social Sciences, *2(12),33-43.*

Introduction

Firstly, from the review of the literature stemmed a confusion about terms, so since in the present research we do not rely on diagnosed behavioral disorders, we consider it reasonable to refer to externalized (behavioral) problems and more specifically to adopt the term of Motti-Stefanidi et al. (2009, p. 255) disruptive and/or aggressive behaviour. The term misconduct, which is mentioned in the Achenbach System of Empirically Based Assessment (ASEBA) in the 1991 version, is considered obsolete, since it is not mentioned anywhere in the later 2003 Greek manual (Achenbach & Rescorla, 2003).

ASEBA serves as a comprehensive framework for evaluating children's and adolescents' skills, adaptive functioning, and behavioral or emotional difficulties. These standardized assessment tools enable the systematic collection of extensive information, supporting consistency, replication, and generalization in the evaluation of behavioral patterns. The main goal of ASEBA is to analyze individual functioning by identifying specific problem areas and empirically derived syndromes using broad-spectrum scales of problems and competencies, supplemented by informant comments. This process generates a rich database that can inform clinical and diagnostic decisions (Achenbach & Rescorla, 2003).

Therefore, ASEBA is an instrument commonly used in school contexts (De Los Reyes et al., 2019) and also a valuable tool in the hands of educators for assessing behavior change and evaluating social validity in SEL interventions. Besides, conducting and publishing a social validity measure of an intervention is crucial, especially when taking into consideration that such publications still remain low, in the range 12-25% (Ferguson, et al., 2018). SEL programs aim to develop skills such as empathy, social awareness, emotional regulation, relationship-building, and responsible decision-making and ASEBA's comprehensive and validated approach makes it well-suited for monitoring the effectiveness of such interventions. A child's level of functioning can differ across contexts and interaction partners, as behavior often depends on situational and relational factors. To capture this variability, diagnostic assessments typically draw on multiple informants—including parents, teachers, and adolescents themselves. This multi-source approach enables a more comprehensive and systematic understanding of children's functioning by allowing comparisons across diverse perspectives (Achenbach & Rescorla 2003 Gresham 2018).

A multi-informant approach offers a valuable means of broadening and enhancing the assessment of individual skills. Gathering evaluations from observers or peers provides a practical, time-efficient, and economically feasible strategy (Stoeffler et al., 2023). Despite these advantages, such assessments may be limited by the possibility that observers' perceptions do not fully capture the true level of competence, and may instead be shaped by egocentric biases (John & Robins, 1993 as mentioned in Stoeffler et al., 2023). Even so, this limitation is balanced by the strength of this method: it relies on externally observed expressions of skills, thereby reducing vulnerability to the self-serving and socially desirable biases commonly found in self-report instruments. Furthermore, research has shown that observer-based evaluations tend to outperform self-reports in predicting various educational outcomes (Wagerman & Funder, 2006).

For all the aforementioned reasons, ASEBA is considered an appropriate evaluation tool for Social and Emotional Learning programs by various researchers in the SEL field (e.g. Elliott & Busse 2004; Haggerty, Elgin & Wooley 2010; Gresham 2018). As far as Special education is concerned, though, the research data about SEL interventions in Greece results from limited research, especially those concerning tertiary intervention programs. Primary prevention programs have been implemented at various periods by the Center for Research and Practical School Psychology of the Kapodistrian University. In the article by Hatzichristou and Lianos (2016) it is mentioned that the primary prevention programs Connecting4Caring Project (2011-2016) and the E.M.E.I.S. (2012-2013) were implemented, but no published results specific to special schools were found. Also, Skeva and Salmond (2015) investigated, through a qualitative research, the attitudes and perceptions of teachers who have implemented SEL programs in their schools, mainly regarding the conditions of implementation in Greek schools, with the aim of highlighting the needs. The results showed that for the more effective integration of SEL programs in Greece, a more active involvement of the school community itself is necessary.

Aim -Objective(s)

The aim of the present article is to highlight the multiple applications of ASEBA in the assessment process of a Social-Emotional Learning program, which was conducted in terms of a mixed-methods research. Our proposition is that ASEBA's applications vary in SEL Interventions, saving valuable time for the researchers/educators, as it may be used as (a) a screening tool and (b) an outcome measure tool as a tool for Social Validity evaluations

As a screening tool, before starting the SEL intervention, ASEBA forms can be administered to establish baseline levels of emotional and behavioral functioning and after the SEL intervention, these forms can be re-administered to measure changes. Measurement of change in intervention studies is challenging, as it involves both accurately measuring a variable at a specific time and tracking how it evolves across several time points (Moreau & Wiebels, 2021). Comparing baseline and follow-up scores helps quantify the intervention's impact on specific behavioral and emotional domains. Also, SEL interventions often target both internalizing behaviors (e.g., anxiety, depression) and externalizing behaviors (e.g., aggression, rule-breaking) and ASEBA's subscales provide detailed insights into changes in both areas. Moreover, ASEBA includes items that assess social competence, such as relationships with peers and social skills, which are directly relevant to SEL goals. In addition, its Multiple Informant Perspective makes it even more important in gathering data from various informants, so as to ensure a comprehensive view of the child's behavior across various settings (home, school, community). This is particularly important in SEL interventions, as changes may be context-specific (Achenbach & Rescorla, 2003).

As a tool for Social Validity evaluations, the ASEBA's multi-informant approach involves parents, teachers, and even students themselves (adolescents), ensuring that the perspectives of those most affected by the intervention are considered. This engagement is crucial for assessing the social validity of an intervention. Also, collecting feedback from these informants about perceived behavioral changes can provide insights into the intervention's acceptability and effectiveness (Gresham, 2018). Moreover, ASEBA offers Norm-Referenced Comparisons by comparing individual scores to normative data, so educators or researchers can determine how students' behavior changes relative to typical developmental expectations, providing a clear context for interpreting results [9]. Last, but not least, as Gresham (2018, p. 119) states, "Using parent and teacher ratings with nationally standardized social skills ratings is one way to quantify the social significance of the intervention's impact. Moving a student's social skills score from the 10th percentile to the 40th percentile would represent a socially significant change.". Similarly, if the targeted behavior shows change, as measured also through direct and systematic observation, and that change falls within the normal range observed among non-referred peers, this would also support the consistency of their behavior rating scale scores and therefore could be considered socially meaningful (Gresham, 2018).

Methodology

Research strategy

Regarding the research methodology, the multiple case study has been chosen as a research strategy, as "it is a research strategy for generalizing to a targeted population of cases through the results of a purposive sampling of cases" (Greene & David, 1984, p. 75). The design of a multiple case study for the analysis and interpretation of the data involves the description of each case separately and then follows a comparison between them to gain an in-depth understanding of the phenomenon under investigation (Yin, 2009). This is the reason why no control group was used; that is, because it would be difficult for 3 more students with the exact characteristics to be found. Therefore, a test-retest design was employed so that change in behavior would be measured within each case and across cases.

Therefore, a purposively selected sample based on diagnosis, age and schooling structure was chosen, in order to investigate the topic and allow for an in-depth investigation and understanding of the issues that the research deals with (Ritchie et al., 2003). Regarding the participants, within the multiple case study for literal replication, 3-4 cases are sufficient, which are similar and the expected results are also similar (Yin, 1994, p. 38-41) and the purposive sampling strategy (purposeful sampling strategy) seems to satisfy the criterion of appropriateness (Patton, 1990 as mentioned in Shakir, 2002).

Moreover, the Single-case experimental design (SCED) with an A-B-A Design was chosen, which includes a return to baseline to verify the intervention effect. According to the literature, ABA studies (two attempts to demonstrate intervention effect) are also accepted as lower-standards Single-case experimental designs (Tate et al., 2013). Cook and Cook (2016) highlight that the most robust evidence regarding whether, and to what extent, interventions lead to improvements in targeted outcomes comes from research employing experimental designs. Single-case experimental designs are a type of experimental methodology used to evaluate the impact of an intervention with a very small number of participants (typically one to three). These designs involve repeated measurements, the sequential (sometimes randomized) introduction of an intervention, and design-specific methods of data analysis, including visual analysis and specialized statistical techniques (Krasny-Pacini & Evans, 2018). Because each individual acts as their own point of comparison, SCEDs are particularly well suited for examining the effects of interventions implemented with individual students or in situations where the sample size is too limited to form both an intervention and a control group (Maggin et al., 2016).

Data was obtained with both qualitative and quantitative methods (mixed method design) in a concurrent triangulation approach, as with the combination of quantitative and qualitative methods the weak points of each method are compensated separately and more substantiated conclusions can be provided through the intersection of findings (Isari & Pourkos, 2015). In this particular paper, though, as already stated, only the assessment process and results of the ASEBA scale are going to be presented.

Participants

The research was implemented in a special education school in the area of Central Macedonia, in Greece. The students were 2 boys (12 years old) and 1 girl (11 years old) and their diagnosis include Mental Retardation of non-organic etiology. They did not a have a diagnosis of Behavioral Problems, but their teachers were stating that they presented severe behavioral problems in class and during the break, difficult to manage. Therefore, choice of students was both criterion-based and purposeful, because they would participate in the research if the results of the ASEBA scale in the Aggressive behavior subscale and the Rule-Breaking behavior sub were at a clinical or borderline level according to most of the informants.

In the assessment process described in this paper 3 teachers and a parent of each of the 3 students participated. George and John were in the same class and Maria in another, so, 3 teachers who were teaching in each class were chosen and questionnaires were also sent home to be filled by both parents, even though only mothers filled it in. It must be noted that the informants from the school context that filled the questionnaires were the teachers who had complained about the classroom management issues they had faced due to the students' behavioral problems, so the sampling was purposeful.

Instrumentation

Achenbach System of Empirically Based Assessment (ASEBA) (Achenbach & Rescorla, 2001) is a toolkit for assessing abilities, adaptive functioning, and behavioral and emotional problems in school-age 6-18-year-olds For the current study, the questionnaires administered were the Child Behavior Checklist-CBCL (Parents' version) and the Teacher's Report Form-TRF (Teachers' version), which both include closed-ended questions of frequency and ratings (0=never, 1=sometimes, 2=often), as well as some open-ended questions, in total 113 questions. Moreover, it is crucial that, although the content of the scales is the same for both genders and for the two age groups 6-11 years and 12-18 years, there are, of course, separate forms for the CBCL and TRF profiles and different norms for each gender and each age group (Achenbach, 2018).

This questionnaire has shown good internal consistency and reliability in previous studies. Within this framework, the ASEBA rating scale was also selected, with its reliability supported by findings from previous studies that have established it as a psychometrically sound instrument. For the CBCL and TRF syndrome scales employed in the present study, Achenbach and Rescorla (2003) reported Cronbach's alpha coefficients ranging from .78 to .97 for the CBCL and .72 to .95 for the TRF, indicating high internal consistency. Regarding cross-informant agreement, the average Pearson correlation coefficient (r) between parent pairs was .76 for the syndrome scales and .60 between teachers.

Procedure-Analysis of data

As stated before, a mixed-methods design was chosen, so quantitative assessment methods were coupled with qualitative during the assessment procedure of the SEL intervention O.M.A.D.A. It must be noted that before the beginning of the SCED, teachers who would implement the Direct Observation procedure, followed a brief practice concerning the familiarization of teachers with methods of Functional Behavior Assessment (FBA). Furthermore, the degree of inter-observer agreement was checked before the evaluation process, as proposed by Gresham et al. (2001).

Therefore, the procedure of the SEL intervention O.M.A.D.A. had the following structure in terms of a Single-Case Experimental Design (SCED):

- ➢ Baseline Phase
- ➢ Design and Implementation of the proposed Social-Emotional Learning Program:
- ➢ Initial assessment and selection of participating students based on the proposed assessment process:
    - o Completion of ASEBA by parents (CBCL version) and teachers (TRF version)
    - o Behavior management strategies: Exploring the views of teachers and parents
    - o Functional Behavior Assessment Interview by teachers, parents and students
    - o Direct observation using the Functional Behavior Assessment Observation Form
- ➢ Design of the individualized intervention program
    - o Diagram of Summary statements and alternative behavior (based on FBA results)
    - o Formulation of a Behavior Support Plan (based on the results of the FBA)
    - o Checking the social validity of objectives and processes (Regarding the Social Validity Check of the objectives, procedures and results of the intervention program before and after the implementation of the intervention program, measurements were made according to an already formulated measurement plan).
- ➢ Intervention Phase
    - o Implementation of the Social-Emotional Learning program.
    - o Formative evaluation through Direct observation using the Functional Behavior Assessment Observation Form
    - o Final evaluation of students with ASEBA

- ➢ Withdrawal phase
- ➢ Evaluation of the SEL intervention program O.M.A.D.A.
    - o Direct observation using the Functional Behavior Assessment Observation Form
- ➢ Follow-up Phase
- ➢ Direct observation using the Functional Behavior Assessment Observation Form
- ➢ Re-evaluation of students with ASEBA
- ➢ Follow-up: social validity measures
    - o Semi-structured interview of parents to check the social validity of the results of the SEL program,
    - o Focus group to record teachers' opinions on the usability of the evaluation tools and the social validity of the results of the SEL program.

The main instrument for evaluation of the students in terms of the SCED design is the direct observation using the Functional Behavior Assessment Observation Form. ASEBA is used mainly for triangulation reasons.

ASEBA, in particular, was administered during the screening procedure, as an initial assessment tool and at the final and follow-up assessments, therefore serving multiple purposes. At the initial assessment, the existence and degree of severity of students' difficulties was ascertained using the ASEBA, so the students then participated in a 9-weeks SEL intervention and then the questionnaires were re-administered to the same informants. After the withdrawal of the intervention for one week, the follow-up assessment took place by administering the questionnaires to the same informants. Therefore, the tool served as a social validity tool as well, since information comes from multi-informants about the perceived change in the everyday life of the students.

As stated in the manual, normalized T-values allow a comparison of a child's position on one scale relative to his position on the others (Achenbach & Rescorla, 2003), therefore, T-values were used in the analysis so as to allow for comparison and conclusions.

The level of agreement between informants Intraclass Correlation Coefficient (ICC) was calculated and specifically the Intraclass Correlation Coefficient ICC(3,1), using a two-factor model with fixed effects of raters and a definition of consistency (two-way mixed, consistency). The statistical analyses were performed using SPSS v24 software (IBM, NY, USA).

Results

Baseline Phase: Initial assessment and Screening procedure results

According to the ASEBA manual, T-values 50-65 indicate the normal spectrum, 66-70 is the borderline spectrum and 70-100 indicate clinical spectrum (Achebnach & Rescorla, 2003).

So, as seen in table 1 below, where there are all the results from all informants for each student, Maria according to her mother has anxiety problems at a clinical level. According to all informants she has Social problems, Rule-breaking Behavior and Aggressive Behavior at a clinical level. Also, according to her teacher, her Physical Education Teacher and her mother she, also, presents Attention problems at a clinical level.

George, first of all, seems to present a different behavior at home from school since his mother, except for Social problems at a low level of borderline, she considers his behavior of normal level. According to his teacher, though, he has anxiety symptoms in an upper level of borderline, all teachers agree on his borderline Rule-breaking Behavior, as well as the clinical level of Social problems, Attention problems and Aggressive Behavior. Last, his Arts teacher considers his thought problems are at borderline level.

As for John, all informants agree on his Rule-breaking and Aggressive behavior at a clinical level, but his teacher seems to be obscured also about Anxiety/Depression, Withdrawal/Depression and Social problems, which she perceives as being at a clinical level. In addition, his Drama teacher, his Arts teacher and his mother agree on his attention problems being at a borderline level. Last, his mother, also, perceives and withdrawal/depression as being at the borderline level.

Table 1. Initial assessment and screening results for Maria, George and Jonh

Syndrome Scale	**Maria**				**George**				**John**			
Informants =>	**T**	**P ht**	**Ct**	**M**	**T**	**Dt**	**At**	**M**	**T**	**Dt**	**At**	**M**
Anxiety/Depression	60	56	58	70	70	51	63	59	72	50	50	54
Withdrawn/Depressed	61	63	56	59	62	58	51	57	78	58	65	68
Somatic Complaints	65	56	61	65	50	50	50	57	64	50	50	57
Social problems	82	82	72	87	70	70	65	66	77	57	64	63
Thought problems	50	62	56	65	71	62	67	63	57	50	62	50
Attention Problems	86	83	67	83	79	76	80	64	65	67	66	67
Rule-breaking Behavior	77	79	79	75	73	73	74	51	71	69	73	71
Aggressive Behavior	86	89	79	94	88	77	95	63	77	71	74	79

****Informants were: Teacher (T), Physical education teacher (Pht), Computers teacher (Ct), Drama teacher (Dt), Arts teacher (At) and Mother (M).***

Pre- and Post- Intervention Comparisons

The mathematical formula used to calculate the percentage of behavior change is: Percentage Change = (Final Score − Initial Score) / Initial Score × 100. Moreover, as mentioned before, according to the ASEBA manual, T-values 50-65 indicate the normal spectrum, 66-70 is the borderline spectrum and 70-100 indicate clinical spectrum. Therefore, the reduction in T-values shows improvement and in the percentage of change is shown with the positive mathematical sign (plus). In contrast, when the T-values are ascending, it means that that a behavior is getting worse, since it is approaching the borderline or the clinical level, and in the percentage of change is shown with the negative mathematical sign (minus). As an Initial score (pre-score) the T-values of the Initial assessment were used, while as a final score (post-score) T-values of the Follow-up assessment we used, so that the overall change and its maintenance would be evaluated.

As far as Maria is concerned (see table 2), there seems to be an overall improvement in her behavior in both contexts, since her teachers and her mother have pointed out great reduction in symptoms in almost all subscales, even though not all of them were targeted in the intervention. The greatest improvement is perceived by her mother (37% of Aggressive behavior), followed by a perceived improvement of attention problems by her teacher (23%) and Physical education teacher's evaluation in the Aggressive behavior (22%). In Aggressive behavior change was, also noted, by the rest of the informants in a smaller percentage (teacher: 12% and Computers teacher: 13%). In addition, the Rule-breaking behavior improved according to her physical education teacher (14%) and her Computers teacher (14%), as well as her mother (11%). Great improvement, though, was noted concerning Attention problems by her teacher (23%), her physical education teacher (20%) and her mother (20%). Last, according to her teacher (15%) and her physical education teacher (17%) her withdrawal/depression symptoms improved, as well as her social problems according to her mother (10%).

It must be noted, though, that there was a minimal negative change in anxiety/depression perceived by her teacher (-5%) and her Computers teacher (-3%), as well as in thought problems stated by her teacher (-3%). It is important to be noted, though, that they still remain in normal level. Except for the aforementioned percentages, all informants rated with positive, though small change the rest of subscales.

Table 2. TRF & CBCL Questionnaires' Results for Maria

Syndrome Scale	**Initial Assessment Pre-score (T-values)**				**Follow-up assessment Post-score (T-values)**				**Change %**			
Informants* =>	**T**	**Pht**	**Ct**	**M**	**T**	**Pht**	**Ct**	**M**	**T**	**Pht**	**Ct**	**M**
Anxiety/Depression	60	56	58	70	63	54	60	70	-5%	4%	-3%	0%
Withdrawn/Depressed	61	63	56	59	52	52	56	52	15%	17%	0%	12%
Somatic Complaints	65	56	61	65	61	50	61	67	6%	11%	0%	-3%
Social problems	82	82	72	87	82	80	70	78	0%	2%	3%	10%
Thought problems	50	62	56	65	52	56	56	61	-4%	10%	0%	0%
Attention Problems	86	83	67	83	66	66	63	66	23%	20%	6%	20%
Rule-breaking Behavior	77	79	79	75	76	68	68	67	6%	14%	14%	11%
Aggressive Behavior	86	89	79	94	76	69	69	59	12%	22%	13%	37%

**** Informants about Maria were: Teacher (T), Physical education teacher (Pht), Computers teacher (Ct) and Mother (M).***

Concerning George, as seen in table 3, the perceived change in, targeted through the intervention, Aggressive behavior subscale was by his Arts teacher (22%), followed by his teacher's ratings (15%) and his Drama teacher (10%). His teacher, also, noticed a change in Rule-breaking Behavior (16%) and in Attention problems (11%). His mother, who in the initial assessment saw no behavior at a clinical level and only worried about Social problems being at a borderline spectrum, still pointed out slight improvement in his behavior in all subscales. What is noteworthy, though, is the increase of anxiety/depression pointed out by Drama teacher (-20%) and his Arts teacher (-5%) and withdrawal/depression symptoms by his Arts teacher (-14%). Controversial seem to be the results concerning social problems, since his Drama teacher and his mother noticed an improvement (9% and (5% accordingly), while his teacher and his Arts teacher noticed a deterioration (-3% and -5% accordingly). It is important to be noted, though, that they still remain in normal level.

Table 3. TRF & CBCL Questionnaires' Results for George

Syndrome Scale	**Initial Assessment Pre-score (T-values)**				**Follow-up assessment Post-score (T-values)**				**Change %**			
Informants* =>	**T**	**Dt**	**At**	**M**	**T**	**Dt**	**At**	**M**	**T**	**Dt**	**At**	**M**
Anxiety/Depression	70	51	63	59	66	61	66	57	6%	-20%	-5%	3%
Withdrawn/Depressed	62	58	51	57	58	55	58	53	6%	5%	-14%	7%
Somatic Complaints	50	50	50	57	50	50	50	54	0%	0%	0%	5%
Social problems	70	70	65	66	72	64	68	63	-3%	9%	-5%	5%
Thought problems	71	62	67	63	65	57	62	58	8%	8%	7%	8%
Attention Problems	79	76	80	64	70	71	78	62	11%	7%	3%	3%
Rule-breaking Behavior	73	73	74	51	61	67	69	51	16%	8%	7%	0%
Aggressive Behavior	88	77	95	63	75	69	74	59	15%	10%	22%	6%

**** Informants about George were: Teacher (T), Drama teacher (Dt), Arts teacher (At) and Mother (M).***

Last, John's results, as shown in table 4, indicate that there is also an overall improvement in his behavior with remarkable improvement in Rule-breaking behavior perceived by his teacher (24%) subscale along with his mother's rating (15%), his Arts teacher's rating (10%) and his Drama teacher's rating (9%). There was, also, a great improvement perceived in the Aggressive behavior by his mother (23%), his teacher (19%), his Arts teacher (16%) and, last, his Drama teacher (14%). Other than the targeted subscales, there was improvement perceived in other subscales, e.g. in Thought Problems, noted by his Arts teacher (19%) and his teacher (12%), even though it was at normal level already in the initial assessment. His teacher also noticed an improvement in subscales: Withdrawal/Depression (15%), Somatic Complaints (11%) and Social problems (12%). His mother also agrees on the improvement in social problems (16%).

Further on, like George, despite the overall improvement, he seems to show a deterioration in anxiety/depression symptoms with his mother's rating (-6%), his Arts teacher's rating (18%) and his Drama teacher's rating (-6%). Also, his Arts teacher noticed an important change in somatic complaints (-28%). It is important to be noted, though, that they still remain in normal level.

Table 4. TRF & CBCL Questionnaires' Results for John

Syndrome Scale	**Initial Assessment Pre-score (T-values)**				**Follow-up assessment Post-score (T-values)**				**Change %**			
Informants* =>	**T**	**Dt**	**At**	**M**	**T**	**Dt**	**At**	**M**	**T**	**Dt**	**At**	**M**
Anxiety/Depression	72	50	50	54	67	53	59	57	7%	-6%	-18%	-6%
Withdrawn/Depressed	78	58	65	68	66	55	62	63	15%	5%	5%	7%
Somatic Complaints	64	50	50	57	57	50	64	57	11%	0%	-28%	0%
Social problems	77	57	64	63	68	57	62	53	12%	0%	3%	16%
Thought problems	57	50	62	50	50	50	50	51	12%	0%	19%	-2%
Attention Problems	65	67	66	67	60	62	59	58	8%	7%	11%	13%
Rule-breaking Behavior	71	69	73	71	54	63	66	60	24%	9%	10%	15%
Aggressive Behavior	77	71	74	79	62	61	62	61	19%	14%	16%	23%

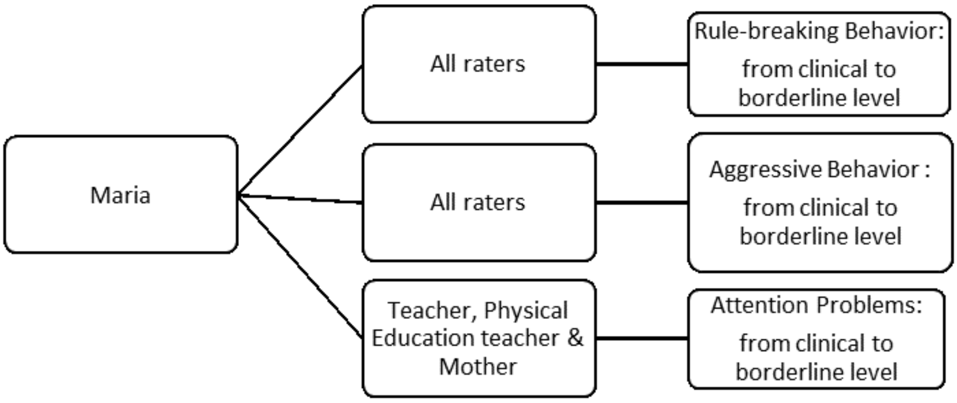
**** Informants about John were: Teacher (T), Drama teacher (Dt), Arts teacher (At) and Mother (M).***

Social Validity evaluation

As aforementioned, according to Gresham (2018, p. 119), "Using parent and teacher ratings with nationally standardized social skills ratings is one way to quantify the social significance of the intervention's impact. Moving a student's social skills score from the 10th percentile to the 40th percentile would represent a socially significant change." Thus, a change in the spectrum (clinical or borderline to normal) of functioning of a child, according to informants, is considered socially important. Reviewing the tables above based on this criterion offers insight into the socially important changes, as perceived by the informants.

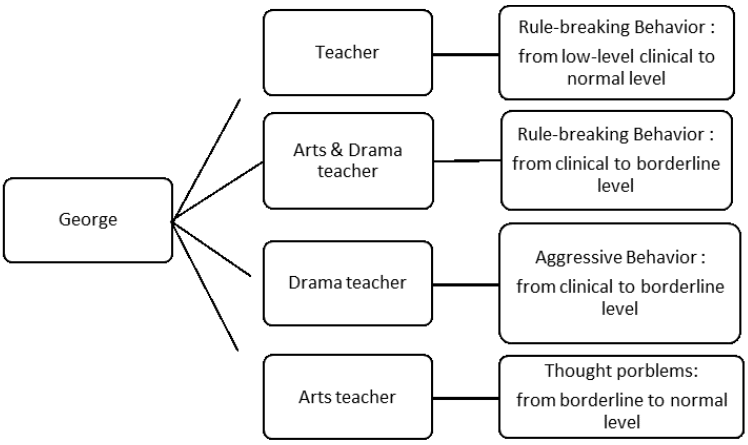
Perceived Change in the spectrum (clinical or borderline to normal) of functioning of the students

Concerning Maria (Scheme 1), there was a significant change in her Aggressive Behavior and Rule-breaking Behavior, according to all informants, from clinical to borderline level. Also, according to her teacher, her Physical Education teacher and her mother a change from clinical to borderline level was noted for Attention Problems as well, even though not targeted through the current intervention.


Maria
All raters
Rule-breaking Behavior: from clinical to borderline level
All raters
Aggressive Behavior : from clinical to borderline level
Teacher, Physical Education teacher & Mother
Attention Problems: from clinical to borderline level

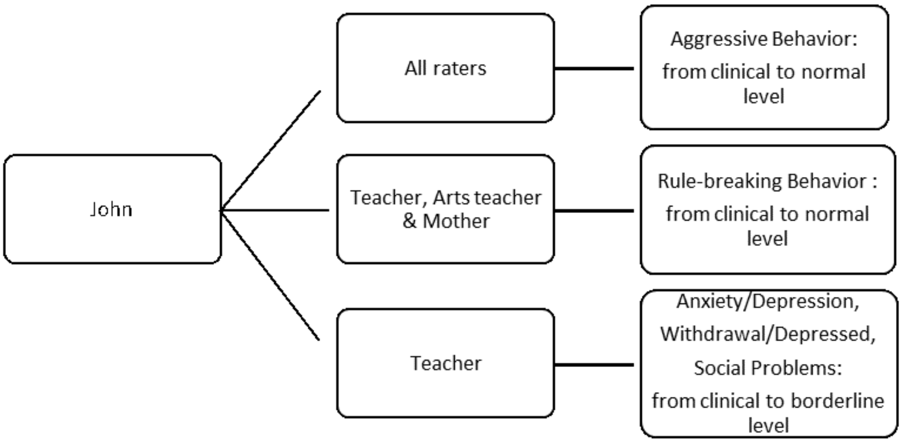
Scheme 1. Perceived Change in spectrum of Maria's functioning

As for George (Scheme 2), his Rule-breaking Behavior changed according to his teacher from low-level clinical to normal level and according to his Drama teacher and Arts teacher from clinical to borderline level. There was also a change in Aggressive Behavior from clinical to borderline level as perceived by his drama teacher. Also, his Arts teacher noticed a change in Thought problems from borderline to normal level


George
Teacher
Rule-breaking Behavior : from low-level clinical to normal level
Arts & Drama teacher
Rule-breaking Behavior : from clinical to borderline level
Drama teacher
Aggressive Behavior : from clinical to borderline level
Arts teacher
Thought porblems: from borderline to normal level

Scheme 2. Perceived Change in spectrum of George's functioning

As far as John is concerned (Scheme 3), a change from clinical to normal level was perceived in Aggressive Behavior by all informants and in Rule-breaking Behavior by his teacher, Arts teacher and his mother. Also, his teacher perceived a change from clinical to borderline level in Anxiety/Depression, Withdrawn/Depressed and Social problems.


John
All raters
Aggressive Behavior: from clinical to normal level
Teacher, Arts teacher & Mother
Rule-breaking Behavior : from clinical to normal level
Teacher
Anxiety/Depression, Withdrawal/Depressed, Social Problems: from clinical to borderline level

Scheme 3. Perceived Change in spectrum of John's functioning

Last, although all three students showed some negative percentage changes, these did not result in a spectrum shift, and all remained at a normal level; thus, the changes were not socially significant, though warranting further investigation.

Intraclass Correlation Coefficient ICC (3,1) of change values

Further on, to investigate the degree of agreement between the four informants who completed the ASEBA questionnaires, the Intraclass Correlation Coefficient (ICC) was calculated. The selection of the appropriate ICC model was based on the guidelines of Koo and Li (2016), who suggest that the choice of model should be based on (a) whether the raters are a random or fixed set, (b) the number of raters, and (c) whether the interest of the analysis concerns absolute agreement or consistency of scoring.

In the present study, in order to examine the degree of agreement between the four raters on the eight ASEBA subscales, the Intraclass Correlation Coefficient ICC(3,1) was calculated, using a two-factor model with fixed effects of raters and a definition of consistency (two-way mixed, consistency), as the four informants were a fixed and specific set and were members of the same interdisciplinary team that systematically participated in the student's assessment process within the context of this multiple case study. Therefore, the purpose of the analysis was to examine the degree of consistency between these specific informants and not to generalize the results to a wider population of potential raters. Therefore, the ICC (3,1) coefficient was used, which is a single-measure reliability index within the context of a two-way mixed model with a definition of consistency. The criteria used were based on the 95% confident interval of the ICC estimate, as proposed by Cicchetti (1994, p. 286) for use in psycholometric assessment results: values less than 0.4, between 0.4 and 0.59, between 0.60 and 0.74, and greater than 0.75 are indicative of poor, fair, good, and excellent reliability, respectively. Specifically, in order to reveal the agreement of raters regarding the perception of the change in the student's behavior, it was considered more important to perform an ICC (3,1) analysis on the change values and not on the T-values, technique mostly used in health-related research studies (e.g. McPhail & Haines, 2010; van Stel et al., 2003). At this point this technique was adopted, since the question needed to be answered was "Do informants agree on how much the student has improved?".

The results indicate poor to fair agreement among raters on the percentage change scores across the ASEBA from baseline assessment to follow-up assessment. Specifically, raters showed fair agreement on Maria's overall change across the whole ASEBA scale (ICC = .591, 95% CI [.244, .880], p= .000), whereas George's raters demonstrated poor agreement (ICC = .251, 95% CI [-.056, .702], p= .062), but with no statistical significance (p= .062 > 0.05). John's raters showed fair agreement (ICC = .466, 95% CI

[.114, .826], p= .003). These findings suggest that the raters differed substantially in how they estimated the degree of improvement of the student across the ASEBA scale, with Maria' s and John's results being statistically significant (p-values <0.05).

As far as the targeted subscales of the externalizing behavior (Aggressive behavior and Rule-Breaking behavior) are concerned, though, the results indicate fair to good agreement among raters. Specifically, informants showed fair agreement on Maria's change of behavior (ICC = .469, 95% CI [.062, .875], p= .010), fair agreement on George's change of behavior (ICC = .567, 95% CI [.157, .907], p= .003), and last, good agreement on John's change of behavior (ICC = .690, 95% CI [.306, .940], p= .000). These findings suggest that the informants agree in how they estimated the degree of improvement across the Aggressive behavior and Rule-Breaking behavior subscales**.** All p-values are <0.05, indicating that the ICCs are statistically significant, thus the agreement among raters is unlikely due to chance.

Discussion

The purpose of the current article was to ascertain ASEBA's multiple applications in SEL Interventions, thus saving valuable time for the researchers/educators, as it may be used as a screening tool, as a baseline assessment, as an outcome/change measure and, also, as a social validity tool.

During the screening process, it helps the researcher examine if students are in the clinical spectrum and which subscales are mostly affected and need a further evaluation or an intervention. In the current study the ASEBA scale did serve its purpose and offered critical information concerning the Screening procedure. All three students were rated by all their informants, except for George' mother, as being at borderline or clinical level in Aggressive behavior and Rule-breaking behavior. Thus, they were all suitable for the participating in the proposed SEL Intervention program. In addition, the ASEBA scores were the Baseline assessment serving as the Pre-score measures in the forthcoming analyses. These results are in accordance with previous research, which proposes that the existence and degree of severity of students' difficulties can be ascertained using the ASEBA psychometric tool, considered by the Collaborative for Academic, Social, and Emotional Learning (CASEL) to be a valuable scale for measuring social and emotional skills (e.g. Gresham, 2018). Of course, the results are not a diagnosis, but can surely serve as a basis for an intervention or as a sign that further evaluation is needed (Achenbach & Rescorla, 2003).

In addition, as found in the baseline assessment, the particular students, even though they had a diagnosis only of intellectual disability, in their everyday life seemed to also show serious signs of Rule-Breaking behavior and Aggressive behavior at a clinical or borderline level, as verified by the ASEBA results. This result is in line with previous research, as it is stated that intellectual disability and behavioral disorders often coexist due to overlapping underlying mechanisms (Einfeld et al., 1996).

Further on, ASEBA was administered again at the Intervention phase and the Follow-up phase, thus serving as as an outcome measure. In that way, the change from the baseline assessment in each scale by each infromant was calculated in percentages. The improvement of deterioration of a student's behavior, as perceived by the informants, was evident, therefore ascertaining ASEBA's designers' notion that reassessment with weighted tools at regular intervals (e.g. every six or twelve months) helps determine the individual's typical course and development (Achenbach & Rescorla, 2003). Therefore, ASEBA can be a valuable tool for measuring an intervention's impact on a student's behavior and, as a result, for deciding if an intervention needs to be continued or not.

Additionally, as far as the evaluation of the outcomes is concerned, a difference of opinions is noted between some informants, which is best explained by the fact that questionnaires like ASEBA fall within the indirect approach of assessment and, as such, they may yield differences as they capture the perceived change, and maybe not the actual change. And in the same way it only serves as a subjective measure of social validity of outcomes. This is why the indirect approaches must be combined with direct approaches, measuring the actual performance of SEL competencies within an authentic context (DiPerna et al., 2023).

Last, as previously stated, ASEBA results come from multi-informants, so in terms of social validity of objectives (initial assessment) and outcomes (perceived change due to intervention) conclusions can be drawn. It was proposed that ASEBA can serve as a social validity measure and it was obvious in this article that it indeed does. A change in the spectrum (clinical or borderline to normal) of functioning of a child indicates a socially valid improvement of the everyday life of the child, as perceived by the informants. In the current research, all informants, except for George' s mother considered the changes in the targeted scales, the Aggressive behavior scale and the Rule-breaking behavior, socially valid, since they indicated a shift in spectrum.

As for as the interrater agreement on the percentage change scores across ASEBA from baseline assessment to follow-up assessment, the results indicate poor to fair agreement among raters. The small number of raters, though, has an impact on the result, since "ICCs are higher if a large number of persons is rated and the population rated reflects the full range of the phenomenon being measured" (Bartko & Carpenter, 1976; Fleiss 1981, as mentioned in Mulsant et al., 2002, p. 1598). Of course, informants' disagreement on their perceived levels of change of a behavior comes as no surprise, since analyzing multiple reports in a single study often results in divergent findings (De Los Reyes, 2013).

Therefore, the differences found, regarding the level of agreement between the informants, might also mean that the objectives of the intervention might not be needed according to some informants, e.g. George's mother. Of course, it could be best explained by the fact that SEL assessments completed by various informants may yield scores with limited validity if the informant has insufficient opportunity to observe the target examinee's SEL competencies over time, or if informants are extremely harsh or lenient in their rating styles (Styck, 2021). Besides, as stated by Kassotakis (1981 as mentioned in Vamvoukas, 2002, p. 313) the evaluation of one person by another is influenced by a multitude of factors related to the personality of both individuals, as the evaluation is often the result of their interaction. Moreover, a possible explanation might be as simple as the different impact of a certain context on the child's behavior, e.g. home or school, especially if significant differences are noted in profiles from different raters (Achenbach & Rescorla, 2003), as in the case of George's mother.

In contrary, the findings concerning the informants' agreement on Aggressive behavior and Rule-Breaking behavior suggest that the informants agree on how they estimated the degree of students' improvement across these subscales**.** The statistical significance of all p-values indicates that the agreement among raters is unlikely due to chance. Overall, John shows the strongest and most reliable agreement for change in the aggressive and rule-breaking behaviors, while Maria's informants show lower and less precise agreement, suggesting more variability among raters concerning the percentage of change. Therefore, the results suggest that John's improvement in aggressive and rule-breaking behavior is perceived more consistently across informants, indicating stronger inter-rater agreement. In contrast, Maria's improvement is evaluated with less consistency, reflecting greater variability among informants in how they perceived the degree of change.

In any case, this comprehensive approach helps in accurately identifying children who may need further assessment or intervention and whose intervention is considered effective according to parents and teachers. Research supports the efficacy of using multiple informants to capture a complete picture of the child's behavior and emotional state (De Los Reyes, 2015).

Of course, certain limitations should be noted. First of all, the smaller number of participating students and their raters may have an impact on the generalizability of findings. The strength of a single study's findings and, thus, the inferences depends on whether different methods of observing, measuring, or analyzing the same behavior consistently yield the same findings (Garner et al. 1956 as mentioned in De Los Reyes, 2013). That is why, even though, ASEBA can be a stand-alone assessment tool, it is better to be a part of a mixed-methods design for the purpose of triangulation, so that ASEBA quantitative results can be verified by qualitative methods, e.g. interviews, direct observation (e.g. Gresham, 2018). In fact, a challenge and a central focus of research and theory on multi-informant assessments is examining the links between ratings from different informants and behavioral data obtained through independent methods, such as naturalistic observations or official records (De Los Reyes et al., 2013).

Conclusion

Nevertheless, ASEBA seems to be a scale easy to administer and score with many important results to be extracted and conclusions to be drawn. It was ascertained, as proposed, that it can offer insight to various, critical aspects of an intervention program, such as screening procedure, baseline assessment, measure of change and social validity.

Concluding, at a first glance, ASEBA seems to be a valuable tool for assessing behavior change and social validity in Social Emotional Learning interventions. Its comprehensive, multi-informant approach along with detailed subscales and normative data, allows for precise measurement of SEL outcomes. By incorporating ASEBA into the evaluation process, educators and researchers can better understand the impact of SEL interventions and make informed decisions to enhance their effectiveness and social validity. In conclusion, despite any limitations, ASEBA surely still remains a valuable tool in the hands of educators who want to implement a SEL intervention in their classroom serving many functions for them and saving valuable time.

Ethics Approval: The program was submitted to the relevant Primary Education Directorate as a Health Education program (Act No. Assignment of the Teachers' Association: 3/9-10-2020).

Data Availability Statement: Data is unavailable due to privacy restrictions, as this paper is a part of a thesis not published yet.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Achenbach, T. M. (2018). Multi-informant and multicultural advances in evidence-based assessment of students' behavioral/emotional/social difficulties. *European Journal of Psychological Assessment, 34*(2), 127–140. https://doi.org/10.1027/1015-5759/a000448
2. Achenbach, T. M., & Rescorla, L. A. (2003). *Enkhirídio yia ta erotimatolóyia kai prophíl skholikís ilikías tou SAEVA* (Epim. Roússou, A.). Elliniká grámmata, Etairía yia tin Psikhikí iyía ton paidión kai ton ephívon.
3. Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284
4. Cook, B. G., & Cook, L. (2016). Research designs and special education research: Different designs address different questions. *Learning Disabilities Research & Practice, 31*, 190–198. https://doi.org/10.1111/ldrp.12110
5. De Los Reyes, A., Cook, C.R., Gresham, F.M., Makol, B.A., & Wang, M. (2019). Informant discrepancies in assessments of psychosocial functioning in school-based services and research: Review and directions for future research. *Journal of School Psychology, 74*, 74-89. doi: 10.1016/j.jsp.2019.05.005
6. De Los Reyes, A., Thomas, S. A., Goodman, K. L., & Kundey, S. M. (2015). Principles underlying the use of multiple informants' reports. *Annual Review of Clinical Psychology, 9*, 123-149. DOI: 10.1146/annurev-clinpsy-050212-185617 (accessed on 5/9/2018)
7. De Los Reyes, A., Thomas, S., Goodman, K., & Kundey, S. (2013). Principles Underlying the Use of Multiple Informants' Reports. *Annual Review of Clinical Psychology, 9*. 123-149. DOI:10.1146/annurev-clinpsy-050212-185617.
8. DiPerna, J.C., Lei, P.W., Anthony, C.J., & Elliott, S.N. (2023). Principles and Practical Approaches to Developing SEL Assessments. In J., Burrus, S., Rikoon, & M.W., Brenneman (Eds.), *Assessing Competencies for Social and Emotional Learning: Conceptualization, Development, and Applications* (pp. 79–98). Routlege. DOI: 10.4324/9781003102243-8
9. Einfeld, S. L., & Tonge, B. J. (1996). Population Prevalence of Psychopathology in Children and Adolescents with Intellectual Disability: II. Epidemiological Findings. *Journal of Intellectual Disabilities Research, 40*(2), 99-109. DOI: 10.1046/j.1365-2788.1996.768768.x
10. Ferguson, J., Cihon, J., Leaf, J., Van Meter, S., McEachin J., & Leaf, R. (2018). *Assessment of social validity trends in the journal of applied behavior analysis, European Journal of Behavior Analysis.* Available online:: https://www.researchgate.net/publication/328304424_Assessment_of_social_validity_trends_in_the_journal_of_applied_behavior_analysis [accessed 29-8-2021].
11. Greene, D. & David, J.L. (1984). A research design for generalizing from multiple case studies, *Evaluation and Program Planning, 7*(1), 73-85. ttps://doi.org/10.1016/0149-7189(84)90027-2.
12. Gresham, F. (2018). Effective Interventions for Social-Emotional Learning. N.Y.: Guilford Publications.
13. Gresham, F. M., Watson, T. S., & Skinner, C. H. (2001). Functional Behavioral Assessment: Principles, Procedures, and Future Directions. *School Psychology Review, 30*, 156-172.

14. Hatzichristou, C., & Lianos, P. (2016). Social and Emotional learning in the Greek educational system: An Ithaca journey. *International Journal of Emotional Education, 8*(2), 105-127.
15. Isari, F. & Pourkos, M. (2015). *Piotikí methodoloyía érevnas: Epharmoyés stin Psikholoyía kai stin Ekpaídefsi*. ISBN: 978-960-603-455-8. Available online: https://repository.kallipos.gr/handle/11419/5826 (accessed on 4/10/2020)
16. Krasny-Pacini & Evans, J. (2018). Single-case experimental designs to assess intervention effectiveness in rehabilitation: A practical guide. *Annals of Physical and Rehabilitation Medicine, 61*, 164–179.
17. Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research*. Journal of Chiropractic Medicine, 15*(2), 155–163.
18. Maggin, D., Cook, B.G., & Cook, L. (2018). Using Single-Case Research Designs to Examine the Effects of Interventions in Special Education. *Learning Disabilities Research and Practice*, *33*(4), 1–10. DOI: 10.1111/ldrp.12184
19. McPhail, S., & Haines, T. (2010). Response shift, recall bias and their effect on measuring change in health-related quality of life amongst older hospital patients. *Health Quality of Life Outcomes, 8*, 65. https://doi.org/10.1186/1477-7525-8-65
20. Moreau, D., & Wiebels, K. (2021). Assessing Change in Intervention Research: The Benefits of Composite Outcomes. *Advances in Methods and Practices in Psychological Science, 4*(1), 1-14. doi:10.1177/2515245920931930
21. Motti –Stefanidi, F., Papathanasiou, A.C., Lardoutsou, S. (2009 Epithetikí kai diataraktikí simperiphorá. In A., Kalantzi-Azizi, & M., Zafeiropoulou, (Eds.), *Prosarmoyí sto skholío: Prolípsi kai antimetópisi*, 9th ed. (pp. 255-286). Elliniká grámmata.
22. Mulsant, B.H., Kastango, K.B., Rosen, J., Stone, R.A., Mazumdar, S., & Pollock, B.G (2002). Interrater Reliability in Clinical Trials of Depressive Disorders. *American Journal of Psychiatry, 159*(9), 1467-1621. https://doi.org/10.1176/appi.ajp.159.9.1598
23. Ritchie, J., Lewis, J., & Elam, G. (2003). Designing and selecting samples. In J., Ritchie, & J., Lewis (Eds.), *Qualitative research practice: A Guide for social science students and researchers*, (pp. 77-108). Sage.
24. Schwartz, I. S., & Baer, D. M. (1991). Social validity assessments: Is current practice state of the art? *Journal of Applied Behavior Analysis, 24*(2), 189-204. DOI: 10.1901/jaba.1991.24-189
25. Shakir, M. (2002). The selection of case studies: Strategies and their applications to IS implementation cases studies. *Research Letters in the Information and Mathematical Sciences, 3*, 69-77.
26. Skeva, A., & Salmont, E. (2015). Epidiṓkontas tēn ensōmátōsē tēs koinōnikḗs kai sunaisthēmatikḗs agōgḗs stis sēmerinés skholikés koinótētes. Stáseis kai táseis pou prokúptoun apó tēn trékhousa ellēnikḗ kai diethnḗ empeiría. In *Proceedings of the Panhellenic Conference of Education Sciences*, 16th June, Athens, Greece, 1299–1306. DOI: 10.12681/edusc.373
27. Stoeffler, K., Rosen,Y. & Way, J. (2023). When Actions Speak Louder than Words: Stealth Assessment of Social and Emotional Skills. In J. Burrus, S.H. Rikoon, & M.W. Brenneman (Eds.). *Assessing Competencies for Social and Emotional Learning: Conceptualization, Development, and Applications*, (p.p. 135-150), Routledge.
28. Styck, K., Anthony, C. J., Flavin, A., Riddle, D., & LaBelle, B. (2021). Are ratings in the eye of the beholder? A tutorial on Many Facet Rasch Measurement to evaluate rater effects in school psychology. *Journal of School Psychology, 86,* 198–221. https://doi.org/10.1016/j.jsp.2021.01.0012021
29. Tate, R.L., Perdices, M., Rosenkoetter, U., Wakim, D., Godbee, K., Togher, L. et al. (2013). Revision of a method quality rating scale for single-case experimental designs and n-of-1 trials: the 15-item Risk of Bias in N-of-1 Trials (RoBiNT) Scale. *Neuropsycholical Rehabilitation, 23*, 619-638. DOI: 10.1080/09602011.2013.824383
30. Tate, R.L., McDonald, S., Perdices, M., Togher, L., Schultz, R., & Savage, S. (2008). Rating the methodological quality of single-subject designs and n-of-1 trials: introducing the Single-Case Experimental Design (SCED) Scale. *Neuropsychological Rehabilitation. 18*(4), 385-401. doi: 10.1080/09602010802009201.
31. Vamvoukas, M. (2002). *Isagoyí stin psikhopaidagoyikí érevna kai methodoloyía*. Grigóris Publications.
32. Van Stel, H.F., Maillé, A.R., Colland, V.T., & Everaerd, W. (2002). Interpretation of change and longitudinal validity of the quality of life for respiratory illness questionnaire in inpatient pulmonary rehabilitation. *Clinical Therapeutics, 24*, 17-18.
33. Wagerman, S. A., & Funder, D. C. (2006). Acquaintance reports of personality and academic achievement: A case for conscientiousness. *Journal of Research in Personality, 41*, 221–229. https://doi.org/10.1016/j.jrp.2006.03.001
34. Yin, R.K. (2009). *Case study research: Design and methods* (4th Ed.). Sage. DOI: https://doi.org/10.33524/cjar.v14i1.73 [accessed on 19/11/2021]
35. Yin, R. (1994). *Case study research: Design and methods* (2nd ed.). 38-41. Sage.