WORKING PAPER Nº IDB-WP-1738

Enhancing Detection of Child Maltreatment and Neglect

An Evaluation of Psychological Tools Adapted for School

María Mercedes Sidders
Guadalupe Soledad Dorna
Santiago M. Perez-Vincent
Norma Peña Arango
María Cecilia López
María Beatriz Müller

Inter-American Development Bank
Institutions for Development Sector
Citizen Security Division

November 2025

IDB

Enhancing Detection of Child Maltreatment and Neglect

An Evaluation of Psychological Tools Adapted for School

María Mercedes Sidders
Guadalupe Soledad Dorna
Santiago M. Perez-Vincent
Norma Peña Arango
María Cecilia López
María Beatriz Müller

Inter-American Development Bank
Institutions for Development Sector
Citizen Security Division

November 2025

Cataloging-in-Publication data provided by the
Inter-American Development Bank
Felipe Herrera Library

Enhancing detection of child maltreatment and neglect: an evaluation of psychological tools adapted for schools / María Mercedes Sidders, Guadalupe Soledad Dorna, Santiago M. Perez-Vincent, Norma Peña Arango, María Cecilia López, María Beatriz Müller.
p. cm. — (IDB Working Paper Series ; 1738)
Includes bibliographical references.
1. Child abuse-Diagnosis-Argentina. 2. Child abuse-Prevention-Argentina. 3. Child abuse-Psychological testing-Argentina. I. Sidders, Mercedes. II. Dorna, Guadalupe Soledad. III. Perez-Vincent, Santiago M. IV. Peña Arango, Norma. V. López, María Cecilia. VI. Müller, María Beatriz. VII. Inter-American Development Bank. Citizen Security Division. VIII. Series.
IDB-WP-1738

JEL Codes: J12, J13
Keywords: child maltreatment, violence against children, early detection, child protection, violence prevention, diagnostic accuracy

http://www.iadb.org

Enhancing Detection of Child Maltreatment and Neglect: An Evaluation of Psychological Tools Adapted for Schools*

María Mercedes Sidders* Guadalupe Soledad Dorna† Santiago M. Perez-Vincent‡
Norma Peña Arango‡ María Cecilia López§,¶ María Beatriz Müller§

Abstract

This study evaluates the diagnostic accuracy of scalable early detection tools that schools can use to identify child maltreatment. Child maltreatment, with severe consequences for children's well-being, remains difficult to detect due to underreporting, stigma, and the limits of existing tools. We conducted a cross-sectional study of over 1,000 students in five schools in Buenos Aires and Córdoba, Argentina, implementing three scalable tools—projective drawings (Kinetic Family and Person in the Rain), the Piers-Harris Self-Concept Scale, and a teacher's checklist—against comprehensive psychological assessments as the reference standard. Results show that approximately 35% of the children in the sample exhibited signs of maltreatment or neglect. Each of the tools demonstrated a precision rate of over 50%, with the combination of projective drawings and teacher's checklist achieving up to 76% precision. The overall sensitivity was 36.8%, significantly enhancing detection rates compared to existing school processes. These findings indicate that early detection tools can significantly enhance identification of maltreatment and hold potential for broader implementation, though they should complement rather than replace comprehensive psychological evaluations.

*The authors thank Catalina Banfi for her invaluable assistance throughout the research process, Franca Genero and Paula Kantor for their expert coordination of the fieldwork, and Sebastián Alarcón for his research assistance. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the Inter-American Development Bank, its Board of Directors, or the countries they represent. This study received approval from the Institutional Review Board (IRB) of the Universidad Abierta InterAmericana in Argentina on March 29, 2023, to conduct the research.

*Fundación Abrazar
†Universidad Torcuato Di Tella
‡Inter-American Development Bank
§Asociación Civil Salud Activa
¶Escuela Argentina de Terapia Simbólico-Proyectiva para el Tratamiento del Abuso Sexual en la Infancia

1 Introduction

The maltreatment of children and adolescents has profound consequences that extend throughout the victim's lives, affecting their physical, emotional, and psychological well-being (Gilbert et al., 2009; Hughes et al., 2017; Mullen et al., 1996; Norman et al., 2012).[1] The World Health Organization (WHO) reports that six in ten children under five—approximately 400 million—regularly experience physical punishment or psychological violence at the hands of parents and caregivers, while one in five women and one in seven men report having been sexually abused during childhood (WHO, 2024). Its high prevalence, combined with its impacts, highlights child maltreatment as a major challenge for social development.

Despite the acknowledged profound consequences of child maltreatment on the physical and psychological well-being of children, effective public policy interventions face numerous challenges. One key issue is the low detection rate, influenced by the complex nature of maltreatment, the limitations of existing detection tools, and substantial underreporting driven by a widespread culture of silence, including social stigma, fear of reprisal, and distrust in child protection systems (Jernbro et al., 2017; Klika et al., 2019; Stoltenborgh et al., 2011, 2013).[2] Recent systematic reviews (Bailhache et al., 2013; McTavish et al., 2020) have further highlighted these detection challenges, particularly the significant limitations of existing child maltreatment detection tools in terms of diagnostic accuracy and suitability for use in general populations.[3] Moreover, confirming cases frequently requires comprehensive psychological evaluations, encompassing a series of tests that demand the presence of trained professionals. These tests are demanding for children and are not easily scalable, typically being used either in legal contexts to build evidence or in consultations when a concerned adult suspects violence toward a child.

The silence and the challenges of deploying detection tools on a large scale make it difficult to fully assess the extent of the problem, which complicates the design and implementation of effective protective measures and interventions. Consequently, evaluating the diagnostic accuracy of simpler, scalable tools becomes crucial for identifying at-risk populations cost-effectively, a necessary step towards more effective public policy in this area.

To address this gap, this study aims to evaluate the diagnostic accuracy of scalable early detection tools within educational settings for identifying child maltreatment and to compare their effectiveness with existing detection practices. We adapted and applied projective drawings (the Kinetic Family and the Person in the Rain) (Hammer, 1958), the Piers-Harris Self-Concept Scale (Piers and Harris, 1964), and a teacher's checklist in a sample of over a thousand students from five schools in the Autonomous City of Buenos Aires and the capital city of Córdoba, Argentina. We then assessed the performance of these

[1] Child maltreatment is the abuse and neglect that occurs to children under 18 years of age. It includes all types of physical and/or emotional ill-treatment, sexual abuse, neglect, negligence and commercial or other exploitation, which results in actual or potential harm to the child's health, survival, development or dignity in the context of a relationship of responsibility, trust or power (WHO, 2024).

[2] This is especially true in the school setting. Interviews with over twenty teachers and school guidance team members revealed that their primary mechanism for identifying at-risk situations is their attentive and close observation. However, they also expressed frustration at not being able to detect severe cases, or cases of children who demonstrate their distress through withdrawal, shyness, and low self-esteem (internalizing behaviors), rather than through externalizing behaviors in the classroom. They cite their daily taskload and the high number of students per class as major contributing factors of this challenge.

[3] For example, McTavish et al., 2020 reviewed 31 studies and found low to very low certainty evidence regarding the accuracy of widely used screening tools, noting high rates of false positives and negatives that could result in both unnecessary interventions and missed cases of maltreatment. Their findings emphasized that most tools were designed for health care or emergency settings, and are ill-suited for mass or selective screening in broader populations such as schools.

tools in identifying cases of maltreatment by comparing their outcomes with those from a comprehensive psychological assessment, which served as the reference test.[4,5]

The comprehensive psychological assessment revealed that approximately 35% of the sample exhibited signs of maltreatment or neglect, underscoring the prevalence of these issues in the educational settings studied. Turning to the diagnostic accuracy of the early detection tools, the tools demonstrated a precision rate (defined as the percentage of red flags that are confirmed as cases of maltreatment) of over 50%, with the combination of projective drawings and the teacher's checklist achieving up to 76% precision. The tools were effective across varying severities and types of maltreatment and socioeconomic backgrounds. However, the presence of a notable fraction of false positives suggests that these tools are better suited for identifying children who may require additional attention or a comprehensive psychological assessment, rather than serving as definitive diagnostic instruments.

Regarding sensitivity, which represents the ability to identify all actual maltreatment cases, we observed that using all three tools together—that is, considering a "red flag" in any of the tools as a positive identification—achieved the highest sensitivity, identifying 36.8% of positive cases. When analyzed individually, the sensitivity rates varied: the checklist identified 13%, the Projective Drawings 25.5%, and the surveys 30% of positive cases.

While this sensitivity rate reveals that most maltreatment cases are not detected by these trio of tools, their application represented a substantial improvement over the current detection standards in the context of our study. In two of the participating schools, we assessed the performance of our tools in contrast to existing detection and intervention processes. We collected data on the number of cases actively being managed by the schools at the time of our assessment. The deployment of our tools resulted in the identification of 38 additional cases of maltreatment, compared to the 22 cases previously recognized by the schools—representing a 173% increase in detection. This underscores the effectiveness of these tools in uncovering new cases beyond the initial identifications made by the school.

While this study shows the significant prevalence of child maltreatment and underscores the effectiveness of diagnostic tools in educational settings for enhancing detection capabilities, detecting maltreatment is just the first step. Effective intervention requires a comprehensive and context-specific approach. We therefore frame this research within a holistic approach to child protection: enhancing early and accurate detection capabilities within educational institutions and engaging the entire school community in proactive prevention and support efforts.

[4]The comprehensive psychological assessment includes several structured components that provide a nuanced understanding of the child's psychological state and converge to classify cases of abuse or neglect. These assessments incorporate techniques such as structured psychological interviews, diagnostic play sessions, and projective techniques (e.g., CAT Test, House-Tree-Person), based on established methodologies (Bleger, 1964a, 1964b; Colombo and Beigbeder de Agosta, 2005; López, 2017). These methods are explained in detail in Section 3.4.

[5]The project involved five critical stages that supported the evaluation of the diagnostic accuracy of the detection tools: 1) obtaining consent from caregivers and children, ensuring ethical compliance and respect for participants' rights; 2) conducting teacher training to familiarize them with the tools and techniques for detecting signs of maltreatment; 3) collecting data using the detection tools; 4) analyzing the data and validating the findings through comparisons with psychological assessments to confirm their precision and sensitivity; and 5) implementing tailored interventions, which, while not directly part of the evaluation, were essential for ethical reasons. These interventions included community-based supportive measures, such as referring families to positive parenting workshops, and more urgent actions, like reporting to child protection agencies when immediate intervention was necessary.

2 Enhancing the Role of Teachers in Child Maltreatment Detection: A Review of Challenges and Tools

This section examines the critical role teachers and schools play in identifying and reporting child maltreatment, the challenges they encounter, and potential solutions offered by specialized risk assessment tools.

Due to their daily interactions with children, teachers are well-positioned to detect signs of maltreatment and serve as key reporters of abuse (Ayling et al., 2020; Benson et al., 2022). However, their effectiveness is often hindered by various obstacles (Dinehart and Kenny, 2015; Pérez de Albéniz Iturriaga et al., 2011), including inadequate training, insufficient knowledge, fear of making false allegations, and potential retaliation from families. These barriers have been documented in studies conducted in various countries, including the United States (Dinehart and Kenny, 2015), Spain (Martinez Sagasti, 2020), Estonia (Toros and Tiirik, 2016), Israel (Haj-Yahia and Attar-Schwartz, 2008), Jordan (Fayez et al., 2014), Sweden (Svensson and Janson, 2008), Taiwan (Feng et al., 2010), and Turkey (Karadag et al., 2015). Further compounding these challenges are normative beliefs that legitimize punitive practices (Feng et al., 2010) and widespread skepticism regarding the efficacy of child protection systems (Choo et al., 2013).

Addressing these barriers requires the implementation of enhanced training initiatives and the deployment of practical, reliable tools (Ayling et al., 2020). Such tools can support teachers in assuming a proactive role in child protection by offering standardized guidelines for recognizing signs of abuse and clear procedures for reporting concerns. By anchoring their assessments in evidence-based protocols, these tools can increase teachers' confidence while minimizing the risk of false reporting or personal repercussions. These issues underscore a persistent gap in school-based child protection: the limited availability of effective, evidence-based instruments for early detection of maltreatment among the general student population.

To assess the current landscape of risk assessment methodologies and their applicability in school settings, the specialized literature was reviewed, with particular attention to their strengths and limitations. Risk assessment tools typically fall into three categories: potential risk tools (which identify risk factors), current risk tools (which detect existing abuse), and recurrence tools (which assess the likelihood of repeated maltreatment). This review focuses on current risk tools, which aim to identify "substantiated cases" of abuse (van der Put et al., 2017).[6]

The majority of these tools are designed for use with families already engaged with child protection systems or within high-risk subpopulations,[7] limiting their utility for proactive identification in broader populations. Among the 18 tools reviewed (van der Put et al., 2017; ICBF, 2022; Engstrom et al., 2020; Algoritmos Públicos, 2021), 83% were developed to support case substantiation in high-risk groups (see Annex 1).[8] This narrow scope excludes a substantial number of children who experience abuse but are

[6] Tools reviewed include those from the systematic review by van der Put et al. (2017) and subsequent studies (ICBF, 2022; Engstrom et al., 2020; Algoritmos Públicos, 2021). The following are the studies included in the systematic review: Sledjeski et al., 2008; Flaherty, 2001; Vaithianathan et al., 2013; Johnson, 2011; Johnson et al., 2015; Dankert and Johnson, 2014; van der Put et al., 2016; Milner et al., 1984; Ayoub and Milner, 1985; Chaffin and Valle, 2003; Ondersma et al., 2005; Coohey et al., 2013; Murphy et al., 1985; Altemeier et al., 1984; Brayden et al., 1993; Wood, 1997; Assink et al., 2015; Camasso and Jagannathan, 1995; Barber et al., 2008; Baird and Wagner, 2000.

[7] For example, tools targeting families with a higher likelihood of qualifying for state benefits, lower socioeconomic status, or contact with the judicial system.

[8] A detailed reference to each of the 18 tools, along with their main characteristics, can be found in Annex 1.

not in contact with formal systems, reinforcing the need for tools that can be deployed universally across school settings.[9]

These limitations are evident both globally and in Latin America and the Caribbean (LAC), the regional focus of the present study. Recent research and expert interviews highlight both challenges and progress in developing risk assessment instruments within LAC (Sidders, 2023).[10] Chile and Colombia have made notable advances, implementing tools such as the Child Alert System (Algoritmos Públicos, 2021) and the Probability Models for Vulnerability Prevention (ICBF, 2022). However, in much of the region, detection continues to rely on informal methods such as teacher observations or sociodemographic interviews by protection teams (Sidders, 2023).[11] These observations highlight the broader regional need for structured, scalable, and context-appropriate instruments.

Although a few tools have been developed for general population use, many rely heavily on administrative data (Flaherty, 2001; Wood, 1997; Murphy et al., 1985). In many countries, such data are often fragmented or unreliable, which presents substantial obstacles to implementation, particularly where data integration systems are weak. Among the three general-population tools identified—the Neural Network Model (Flaherty, 2001), NCCD Risk Assessment Tools (Wood, 1997), and the Family Stress Checklist (Murphy et al., 1985)—the first two depend extensively on administrative sources. None of these tools were originally designed for use in Latin America, and adapting them can be as resource-intensive as developing new tools from scratch.

Overall, the current landscape reveals several critical gaps. Most tools are limited in scope to populations already involved with protection systems and are dependent on administrative data, which reduces their applicability for early identification in the general population. Tools that do target broader populations frequently lack sufficient contextual adaptation, hindering their feasibility in diverse and low-resource environments. These constraints underscore the need for tools that are user-friendly, scalable, and adaptable, particularly by non-specialist personnel such as teachers. Addressing these challenges involves not only refining technical aspects but also ensuring that tools are viable for routine use in school settings. Based on the literature, the following features are proposed as essential for effective risk assessment tools:

Child-Safe Participation: Tools should facilitate non-intrusive ways for children to share their experiences. Techniques such as drawings and self-concept surveys are recommended, particularly for children who may be living with abusive caregivers. These indirect methods help prevent re-traumatization and offer safer avenues for disclosure (López, 2015; Martínez, 2014; Serrano and Villalba, 2018).

Addressing Systemic Biases: Many tools carry biases inherent in administrative datasets, often disadvantaging economically vulnerable families (Center and Hall, 2018; Pryce et al., 2019). Tools should incorporate psychological and behavioral indicators that reflect diverse cultural and socioeconomic contexts to mitigate these effects.

[9] Additionally, McTavish et al., 2020 reviewed 31 studies assessing the diagnostic accuracy of tools primarily designed for use in healthcare and emergency settings. These tools, which include decision rules, checklists, and physical abuse protocols, further illustrate the predominance of instruments intended for high-risk or system-involved populations. Their findings reinforce the need for scalable tools appropriate for universal deployment in schools, where early signs of maltreatment may first surface.

[10] Sidders, 2023 conducted semi-structured interviews with 30 experts and officials from across the region. Interviewees were selected to ensure diversity in regional and sectoral representation, including child protection, education, health, and judiciary sectors. The sample included participants from Argentina, Mexico, Chile, Colombia, Brazil, Honduras, Paraguay, Peru, Ecuador, and Bolivia.

[11] The Chilean and Colombian tools use administrative data to predict substantiated maltreatment cases within selected subpopulations.

Accessibility for Non-Specialists: Tools should require minimal specialized training and be straightforward to use in varied educational environments. Instruments should be designed for effective application by educators and other school personnel without relying on child protection or judicial experts. In the reviewed literature, 53% of tools required specialized personnel, whereas only 18% were suitable for broader frontline use.

Ensuring Political Acceptability: The sustainability of tool adoption depends on community and institutional acceptance. Tools should be framed in non-accusatory terms, promoting trust and clear procedural guidance to encourage use by school staff and administrators.

3 Diagnostic Tools and Reference Test

Building on the gaps identified in the literature and the desired features for risk assessment tools, this study evaluates the diagnostic accuracy of innovative early detection tools within educational settings. Schools, particularly in contexts where primary school attendance rates are close to 100%, such as Argentina, provide an ideal environment for implementing tools that proactively identify at-risk children in a broad, inclusive manner.

In this study, we tested three diagnostic tools adapted for educational settings, drawing from both psychotherapy practices and broader screening methodologies. These tools were selected to align with the previously discussed features by being accessible for non-specialist use by school staff,[12] incorporating non-intrusive approaches to identify potential maltreatment, and ensuring practical applicability.

The early detection tools included projective drawings, the Piers-Harris Self-Concept Scale, and a teacher's checklist. For each tool, we established specific criteria to trigger a positive case indication, or a "red flag," signaling potential maltreatment or neglect. The performance of these tools was assessed by comparing their outcomes with those from a comprehensive psychological assessment that integrated indicators from five distinct psychological techniques (Psychological Interview with Children, Diagnostic Play Hour, Children's Apperception Test (CAT), Graphic Projective Technique (House-Tree-Person, H.T.P.), and Desiderative Questionnaire), which served as the reference test.

3.1 *Projective Drawings (Annex 2)*

The projective drawing tool comprises two specific exercises: the "Person in the Rain" exercise and the "Kinetic Family Drawing" exercise. In the "Person in the Rain" exercise, children are asked to draw a person standing in the rain, while in the "Kinetic Family Drawing" exercise they draw all of their family members (including themselves) engaged in an activity.

Analysis of the drawings centers on identifying graphic indicators that may reflect a child's emotional state or experiences. Deriving from an extensive body of research, we compiled a set of such indicators that are traditionally examined when interpreting the Person in the Rain and Kinetic Family drawings.

[12] Although the proposed tools allow for broader data collection, the analysis phase, especially for drawings, still requires child psychology expertise.

This approach is founded on the work of early pioneers such as Paula Elkisch (1966), Luis Corman (1967), and Marisa Rodulfo (1992), who initially explored the psychological insights that children's drawings can offer. Elizabeth Münsterberg Koppitz (2006) later refined this approach by introducing developmental indicators into human figure drawing assessments to improve screening and diagnostic utility. Müller and López (2011) were among the first to examine multiple drawing tests jointly to identify common graphic indicators of abuse or maltreatment. Subsequent researchers, including Chong (1987) and León Vásquez and Castañeda Chang (2012), applied these methods to identify emotional disturbances in children with a history of abuse. Colombo et al. (2013) specifically focused on the Person in the Rain test, employing a rigorous statistical approach. They conducted three rounds of analysis comparing drawings from children with confirmed abuse cases to those from the general population, calculating the frequency of each graphic indicator in both groups. This process advanced the standardization of the Person in the Rain test.[13]

Building on this prior research, we identified 24 pre-selected indicators for the Person in the Rain test and 50 for the Kinetic Family test (see Annex 2 for the complete list). These indicators include specific graphic elements such as missing body parts, phallic symbols, figures drawn without clothing, and amorphous shapes, which are among the features traditionally examined when interpreting these projective drawings.

For the criterion to trigger a "red flag," psychologists examined each child's drawings for the presence of these pre-selected indicators and then rated each drawing's overall risk level (low, medium, or high). A case was flagged only if the Person in the Rain and Kinetic Family drawings were rated as high risk by the evaluator. This dual high-risk criterion was pre-specified in light of prior reviews cautioning against over-interpretation of single projective signs (Allen and Tussey (2012); Garb et al. (2000)).

3.2 *Piers-Harris Self-Concept Scale (Annex 3)*

The Piers-Harris Self-Concept Scale (Piers and Harris, 1964) was chosen to assess self-perception and identify behaviors that may suggest underlying experiences of violence, such as withdrawal, shyness, low self-esteem, and aggression. This scale consists of 80 items requiring "Yes" or "No" responses to statements like "I am a happy person" and "I find it hard to make friends" (see Annex 3 for the complete list). It covers six content domains: behavioral adjustment, intellectual and school status, physical appearance and attributes, freedom from anxiety, popularity, happiness and satisfaction (Piers and Herzberg, 2002), providing a comprehensive view of a child's self-esteem and general self-view.

The potential value of this instrument for identifying child maltreatment relies on a large body of research linking childhood trauma and maltreatment with a more negative or diminished self-concept. Meta-analytic evidence indicates a robust association between maltreatment exposure and lower self-concept or self-esteem in children and adolescents, with stronger effects for cumulative or repeated trauma and for certain subtypes such as sexual abuse (Melamed et al., 2024; Zhang et al., 2023). Comparative

[13] Evidence on the *diagnostic accuracy* of projective drawings remains mixed. Two critical syntheses conclude that, given the available controlled research, no single drawing indicator or scoring system demonstrates sufficient validity to diagnose abuse on its own; positive findings are rarely replicated and many studies have important methodological limitations (Allen and Tussey (2012); Garb et al. (2000)). A broader systematic review likewise judged the evidence inconclusive, citing heterogeneous designs, case–control sampling, and the absence of a consistent reference standard, while noting some group-level differences that warrant further study (Veltman and Browne (2002)). More recent large-sample studies report drawing patterns that are more prevalent among victimized youth, which supports the plausibility of projective drawings as *risk markers* within multi-method assessments; however, rigorous diagnostic-accuracy evidence (e.g., against blinded reference standards with prespecified thresholds) is still limited (Jaroenkajornkij et al. (2022); Ballús et al. (2023)).

studies using the Piers–Harris specifically have found that abused children show lower self-concept than matched non-abused peers (Oates et al., 1985). Related work in traumatized samples shows depressed Piers–Harris scores among youth with post-traumatic symptomatology compared to non-traumatized controls (Saigh et al., 2008). Additionally, the Piers–Harris total score typically demonstrates high internal consistency (often around .90) and acceptable temporal stability in child samples, including modern computer-administered formats, supporting reliable group administration and interpretation (Flahive et al., 2015). Taken together, these findings support the instrument's potential to surface maltreatment-related distress at the group level, justifying a formal test of diagnostic accuracy in school settings.

To score and standardize, first, items were reverse-coded where required and summed by domain.[14] The total score was obtained as the sum of the domains. To ensure comparability across administrations, we expressed the total score as a percentile within the study's reference groups (three schools in Buenos Aires in the first round, followed by one school in Córdoba and then another school in Córdoba).

A red flag for "diminished self-concept" was defined *a priori* as a total score at or below the 15th percentile within the relevant reference group. Traditionally, a good self-concept is indicated by scores above the 65th percentile, while a poor or diminished self-concept is flagged below the 35th percentile. This more stringent threshold was adopted based on the understanding that severe cases of maltreatment and neglect are more likely to manifest through markedly diminished self-concept, necessitating a stricter standard for early identification.[15]

3.3 *Teachers' Checklist (Annex 4)*

The third tool implemented in our study is the teacher's checklist, which consists of 20 specifically designed indicators to help teachers observe and record signs of child abuse or neglect within the school setting (see Annex 4 for the complete Teacher's Checklist). These indicators were selected based on protocols and guidelines from national and provincial governments in Argentina, including those by Averbuj (2010); Bianco et al. (2016); Gobierno de la Ciudad Autónoma de Buenos Aires (2021); Gobierno de la Provincia de Buenos Aires (2019, 2022); and Gobierno de la Provincia de La Pampa (2017). All indicators from these documents were compiled and prioritized based on their relevance and ease of identification for teachers. Conceptually, teacher observation tools leverage teachers' day-to-day proximity to children and their ability to notice persistent patterns across settings (classroom, playground, arrival/dismissal), which can capture both externalizing and internalizing manifestations of maltreatment that may not be evident in single-time clinical encounters. The checklist includes indicators such as physical signs (e.g., frequent injuries), behavioral changes (e.g., explicit sexual behavior), academic issues (e.g., frequent absences), and psychological indicators (e.g., reports of maltreatment).

Evidence supporting the potential utility of structured teacher observations comes from both psychometric studies of classroom checklists designed to flag trauma-related behavior and studies of teachers' detection and reporting performance. Leeuwestein et al. (2024) examined the properties of a teacher observation instrument for early primary students, which demonstrated coherent factor structure and

[14] Items 7 and 12 belonged to two dimensions simultaneously, so they were counted twice.

[15] We provide a threshold analysis in Annex 8. This analysis demonstrates how the performance of our detection tools changes as we vary the threshold for triggering a red flag, illustrating the sensitivity of our results to different threshold choices.

acceptable internal consistency, indicating teachers can reliably rate behavior clusters associated with psychotrauma when given structured indicators. Evidence shows that strengthening school detection and reporting pathways increases case surfacing without reducing report validity: a systematic review finds that child-protection training improves professionals' reporting behaviors (although evidence is characterized as low-certainty (Walsh et al. (2022)), and a quasi-experimental analysis of U.S. state mandates for school-based prevention education found higher rates of child sexual abuse reports by education personnel with no decrease in confirmation rates (Bright et al. (2022)).

The criteria for triggering a red flag on the checklist are as follows: Teachers complete the checklist by marking "Yes" or "No" for each indicator based on their observations of the student's behavior and appearance. The presence of any physical indicator alone, such as frequent unexplained injuries, or the presence of more than one behavioral or academic indicator, is sufficient to classify a case as high-risk. Additionally, any direct disclosure of abuse by the child is an automatic trigger for a red flag.[16]

3.4 *Reference Test: Comprehensive Psychological Assessment*

To assess the diagnostic accuracy of the early detection tools described above, we selected a structured comprehensive psychological assessment to measure concurrent validity. This choice follows widely endorsed principles for child-protection evaluations: using a multi-method approach that integrates interview, behavioral observation, and standardized techniques to reduce the risk of both false positives and false negatives (American Professional Society on the Abuse of Children (APSAC) (2022), American Psychological Association (2024), and National Institute for Health and Care Excellence (NICE) (2017)). The rationale for selecting this method as the reference standard was its ability to provide a nuanced, evidence-based assessment of maltreatment while minimizing false positives and avoiding revictimization, particularly in sensitive cases.

This test is used to determine whether the "red flags" raised by the proposed tools are confirmed as cases of maltreatment or neglect, and vice versa. In a small number of cases (n=13), where prior confirmation of maltreatment was available from school records or reports, the psychological assessment was not conducted in order to minimize the risk of revictimization through redundant evaluation. These cases are counted as confirmed cases of maltreatment for our analysis. This procedure is consistent with guidance to minimize repetitive evaluations and to prioritize child welfare when adequate corroboration already exists (National Institute for Health and Care Excellence (NICE), 2017).

The comprehensive psychological assessment includes several components, which provide a deep and nuanced understanding of the children's psychological state and their experiences of maltreatment. However, important limits remain: many abused children (particularly in cases of sexual abuse) do not disclose during any single evaluation, and disclosure likelihood depends on interview quality, child age, relationship to the perpetrator, and context (Azzopardi et al., 2019; Fernandes and colleagues, 2024; Lavoie et al., 2021). Interviewing protocols can improve interviewer behavior and children's informativeness by increasing open invitations and reducing suggestive prompts (Benia et al., 2015), but even under optimal conditions a meaningful fraction of victims will not disclose (Azzopardi et al., 2019). To account for these dynamics, the assessment integrates both direct testimony and indirect indicators drawn from projective and observational techniques, allowing for a more comprehensive and age-sensitive

[16] See Annex 8 for threshold analysis.

identification of abuse and neglect. Still, while the assessment captures a broad range of cases, some experiences—particularly those involving sexual abuse—may be underidentified.

To classify cases as positive for abuse or neglect, the comprehensive psychological assessment required convergence across methods. A positive classification required the convergence of indicators from the following components:

1. **Psychological Interview with Children:** This involves a concise, structured conversation to elicit information from the child about their experiences. The interviewer explores the child's responses and any issues or questions that arise, providing a direct avenue for identifying verbal indicators of abuse or neglect. This approach is based on a technique developed by Bleger (1964a, 1964b) and later adapted by López (2017) to address the specific needs of interviewing children who are victims of abuse, which includes evidence-informed practices (non-leading, open questions, rapport building) consistent with forensic interviewing guidance (American Professional Society on the Abuse of Children (APSAC), 2022; Benia et al., 2015; Lavoie et al., 2021).

2. **Diagnostic Play Hour:** Observation focuses on the child's behavior, interaction with play toys and material, and verbal expressions during a 45-minute play session. Indicators of maltreatment may emerge through how the child relates to the play scenario, their choice of words, and the emotions they exhibit. The technique we use originates from Freud (1920) psychoanalytic theory of pleasure and Aberastury (1962) projections, and was later adapted by Colombo and Beigbeder de Agosta (2005) to focus on identifying signs of abuse and maltreatment in children. López (2014) further expanded on the role of play in detecting sexual abuse. Consistent with guidelines, play-based observations were interpreted as clinically informative but not determinative in isolation (American Professional Society on the Abuse of Children (APSAC), 2022; National Institute for Health and Care Excellence (NICE), 2017).

3. **Children's Apperception Test (CAT):** Uses the projection of feelings and conflicts onto drawings to assess a child's perceptual and expressive abilities, personality, and relational conflicts. Although the CAT was not specifically designed to detect abuse, various authors have reviewed its utility in identifying indicators related to maltreatment. Sánchez (2012), Gómez (2018), Colombo and Beigbeder de Agosta (2005), and López (2014) have explored how certain patterns, themes, and emotional responses that emerge in CAT narratives can suggest experiences of abuse or neglect, and how professionals can interpret these signs in clinical settings. In our reference standard, CAT themes contributed corroborative weight only when consistent with other components (American Professional Society on the Abuse of Children (APSAC), 2022).

4. **Graphic Projective Technique (House–Tree–Person, H.T.P.):** This technique captures involuntary and unconscious projections of the self onto paper. Indicators of maltreatment are discerned through the child's depiction of objects, people, and scenarios, with particular attention to the attributes, feelings, and efforts projected, offering a window into their self-perception and emotional state. Blain et al. (1981) and Morín Díaz and Ramírez Figallo (2017) reviewed how patterns in HTP drawings can reveal experiences of abuse or trauma, providing guidelines for interpreting key indicators of child maltreatment. As with the CAT, HTP findings were treated as supportive when convergent with interview and observational evidence, in line with cautions from practice guidelines (American Professional Society on the Abuse of Children (APSAC), 2022; American Psychological Association, 2024).

5. **Desiderative Questionnaire:** By asking the child to imagine themselves as a non-human entity, this technique delves into the child's fantasies, anxieties, and defenses. Maltreatment indicators may surface through the child's choices and the reasons behind them, revealing aspects of their self-esteem, sexual identity, and personal values, as well as enabling the articulation of memories and associations that may relate to experiences of abuse or neglect. Gómez (2015) and Hernández (2019) explored how the Desiderative Questionnaire can identify trauma in children, while Sánchez (2020) and López (2020) analyzed its use in detecting sexual abuse, offering interpretation guidelines. As with other projective methods, these results informed the overall synthesis but were not used as stand-alone proof (American Professional Society on the Abuse of Children (APSAC), 2022; American Psychological Association, 2024).

In sum, a positive case for abuse or neglect is determined by the convergence of positive indicators from these techniques, underpinned by the child's own testimony about their violent experiences. This convergence standard (multi-method agreement over single-test signals) is consistent with international guidance emphasizing triangulation and careful differential diagnosis in suspected maltreatment (American Professional Society on the Abuse of Children (APSAC), 2022; American Psychological Association, 2024; National Institute for Health and Care Excellence (NICE), 2017). The assessment also allowed us to classify positive cases in our study into three distinct categories based on the severity of the situation and the frequency of violent incidents: urgent, severe, or moderate. Urgent cases were those with a potential risk of death or serious harm, typically involving physical and sexual abuse or extreme negligence. Severe cases, while not posing an immediate threat to life or causing serious harm, involved recurrent violence, occurring at least twice a year or more. Moderate cases were characterized by sporadic incidents of violence, happening no more than once a year. Examples of severe and moderate cases include repeated instances of neglect, inadequate supervision, and verbal or physical aggression. The positive cases were also categorized based on the type of abuse experienced. This included physical abuse, psychological abuse, neglect, sexual abuse, and exposure to violence. Despite a comprehensive protocol, misclassification remains possible—particularly for sexual abuse without disclosure—so the diagnostic accuracy of the index tests should be interpreted in light of this imperfect but evidence-based reference standard (Azzopardi et al., 2019; Lavoie et al., 2021; National Institute for Health and Care Excellence (NICE), 2017).

4 Implementation and Validation Process for Diagnostic Tools

We conducted the assessment of the three diagnostic tools across five private primary schools in Buenos Aires City and the capital city of Córdoba, Argentina. A formal agreement was established between the research team and each participating school, detailing procedures and timelines. The implementation took place in three phases throughout 2023, allowing continuous refinement and enhancement of our strategies. Phase one, from April to October, included three schools in Buenos Aires City. Phases two and three, from July to September and September to November respectively, focused on different schools in Córdoba.

The schools in Buenos Aires were recently established and had been gradually expanding by adding new grade levels each year. At the time of the study, one school covered grades 1 to 7, another covered grades 1 to 6, and the third covered grades 1 to 4. The schools in Córdoba followed the standard primary school

structure, serving grades 1 to 6. All families in these schools were invited to participate in the study.

The socioeconomic context of the areas served by these schools varied: three were located in low socioeconomic areas, one in a lower-middle socioeconomic area, and the fifth in a middle socioeconomic area. Despite being private, the schools in low socioeconomic areas did not charge fees, the lower-middle socioeconomic school charged minimal fees that were often waived for families in need, and the middle socioeconomic school charged standard fees. This diversity allowed us to test the diagnostic tools across varied socio-demographic settings.

Our validation process consisted of four stages, which were integral to assessing the diagnostic accuracy of the tools: obtaining consent from caregivers and children, training teachers, collecting and analyzing data, and validating results through comprehensive psychological assessments.

This was a prospective diagnostic accuracy study, with data collection and procedures defined prior to the application of the index tests and the reference standard.[17] The eligibility criteria for the screening phase included all students enrolled in grades 1 to 7 at the five participating schools, provided informed consent was given by their caregivers and assent by the children. For the reference assessment phase, children identified by the screening tools were excluded only if they had cognitive or communication disabilities that precluded the valid administration of a comprehensive psychological evaluation.

Each stage is outlined below:

Stage 1: Consent. The initial phase of our study consisted in obtaining informed consent from primary caregivers, both parents or guardians, and subsequently the children themselves. We prepared detailed information sheets and consent forms[18], which were distributed to families through the school staff. These documents outlined the main aspects of the research, ensuring that caregivers were sufficiently informed before giving their consent. Additionally, we also asked for children's assent, seeking their oral consent directly to participate in the research activities.

The target population for our study was 1,365 children from the selected schools. Of these, 1,137 families and children consented (83.3%). Figure 1, below, presents the key figures at each stage of the study.

Stage 2: Application of Diagnostic Tools and Teacher Training. In this stage, data collection was conducted using the three diagnostic tools described previously: projective drawings and the Piers-Harris Self-Concept Scale, completed by the children, and a comprehensive teacher's checklist of 20 indicators,

[17] Beyond these four stages, the project involved supporting schools to implement tailored interventions, which was done to provide necessary support to children identified as needing assistance. The study received formal approval from the Ethics Committee for Scientific and Technological Research of the Universidad Abierta Interamericana on March 29, 2023.

[18] We implemented an informed active consent process, particularly tailored to the specific objectives of our project. Crucially, we chose not to detail the specific benefits of the research in the consent documentation. This decision was based on our belief that requesting full consent might not be a reasonable requirement for protecting the subjects, in this case, children who have experienced violence. Our study met the three conditions often required for such choice: (a) the research involves no more than minimal risk to the participants; (b) the waiver will not adversely affect the rights and welfare of the children; and (c) the research could not practicably be carried out without the modification. Consequently, we sought parental consent for the children's participation in the activities, with the consent form communicating the general purpose of the project in a positive manner, phrasing the general purpose as improving children's wellness. This decision was made to prevent parents, guardians, or caregivers who might be perpetrating violence from withdrawing the children from the study. Additionally, as previously mentioned, our tools, including observational methods like projective drawings and a non-sensitive survey, are non-intrusive. Therefore, the consent was carefully crafted to avoid making parents, guardians, or caregivers feel accused or defensive. This approach was integral to ensuring the participation of all children, including those potentially at risk, without alerting their caregivers.

used by teachers to observe signs of child maltreatment and neglect. Teachers were trained on the teacher's checklist by staff of the Fundación Abrazar. The training encompassed equipping educators with the ability to identify signs of violence, engage sensitively with victims, and implement preventative strategies within their classrooms.

The Projective Drawings tool was administered to children across all grade levels. The Piers-Harris Self-Concept Scale was administered only to older children, specifically those in 5th, 6th, and 7th grades, given its complexity and the reading proficiency required. In the case of the teacher's checklist, in the first round of schools, teachers were asked to complete it only for students who showed red flag in their drawings or survey responses. However, in subsequent rounds, we expanded the teacher's checklist's application, asking teachers to observe all children in their classrooms.

Out of the consenting 1,137 children, we collected data from 1,087 using at least one diagnostic tool (the Projective Drawings), with some attrition due to absences. Table A.2 reports sex and age distribution of the children in the sample. 370 children (5th to 7th graders) completed the Piers-Harris Self-Concept Scale, and the teacher's checklist was completed for 630 children. We obtained data from all three tools for 223 children. No indeterminate results were reported for the index tests. Cases with missing data on the index tests were excluded from the diagnostic accuracy analysis; no imputation procedures were applied. No adverse events or distress requiring clinical response were reported during the administration of the screening tools.

Stage 3: Analysis of Diagnostic Tools. The analysis involved psychologists evaluating drawings, calculating Piers-Harris Self-Concept Scale scores, and assessing checklist indicators. All assessments were conducted under blinded conditions to prevent bias arising from prior knowledge of students' backgrounds or results from other tools.

Regarding the drawing assessments, psychologists received training from Fundación Salud Activa on how to identify indicators in projective techniques. The entire psychology team, supervisors, and specialists from Fundación Abrazar participated. Weekly group supervisions provided ongoing support, focusing on indicator identification. The evaluation process was blind, and supervisors independently analyzed 13% of the sample. The reliability of the ratings between supervisors and psychologists was calculated using Cohen's Kappa. The observed agreement was approximately 0.896, indicating that the raters agreed on 89.6% of the evaluations. The expected agreement by chance was approximately 0.420. The resulting Kappa value was 0.82, signifying a near-perfect agreement between the supervisors and psychologists. This analysis was conducted at the aggregated level, considering the child's ID rather than each individual drawing, meaning the evaluation took into account both drawings per child. This high Kappa value demonstrates strong reliability in the classification of the drawings.

High-risk cases were identified using these tools according to the criteria described in the previous section, with children anonymously flagged for further psychological assessment.

From the diagnostic tools, 278 children (25.3%) were flagged as at risk. In line with the ethics protocol, all of these children advanced to the psychological evaluation phase. Additionally, 57 children without a red flag were randomly selected for psychological assessment to estimate the overall prevalence of maltreatment and evaluate false negative rates. In total, 335 children proceeded to the validation stage, during which comprehensive assessments were conducted. Sample sizes were determined based on

the statistical power required to assess the precision of the diagnostic tools (see Annex 5), while also minimizing unnecessary exposure of children to comprehensive assessments.

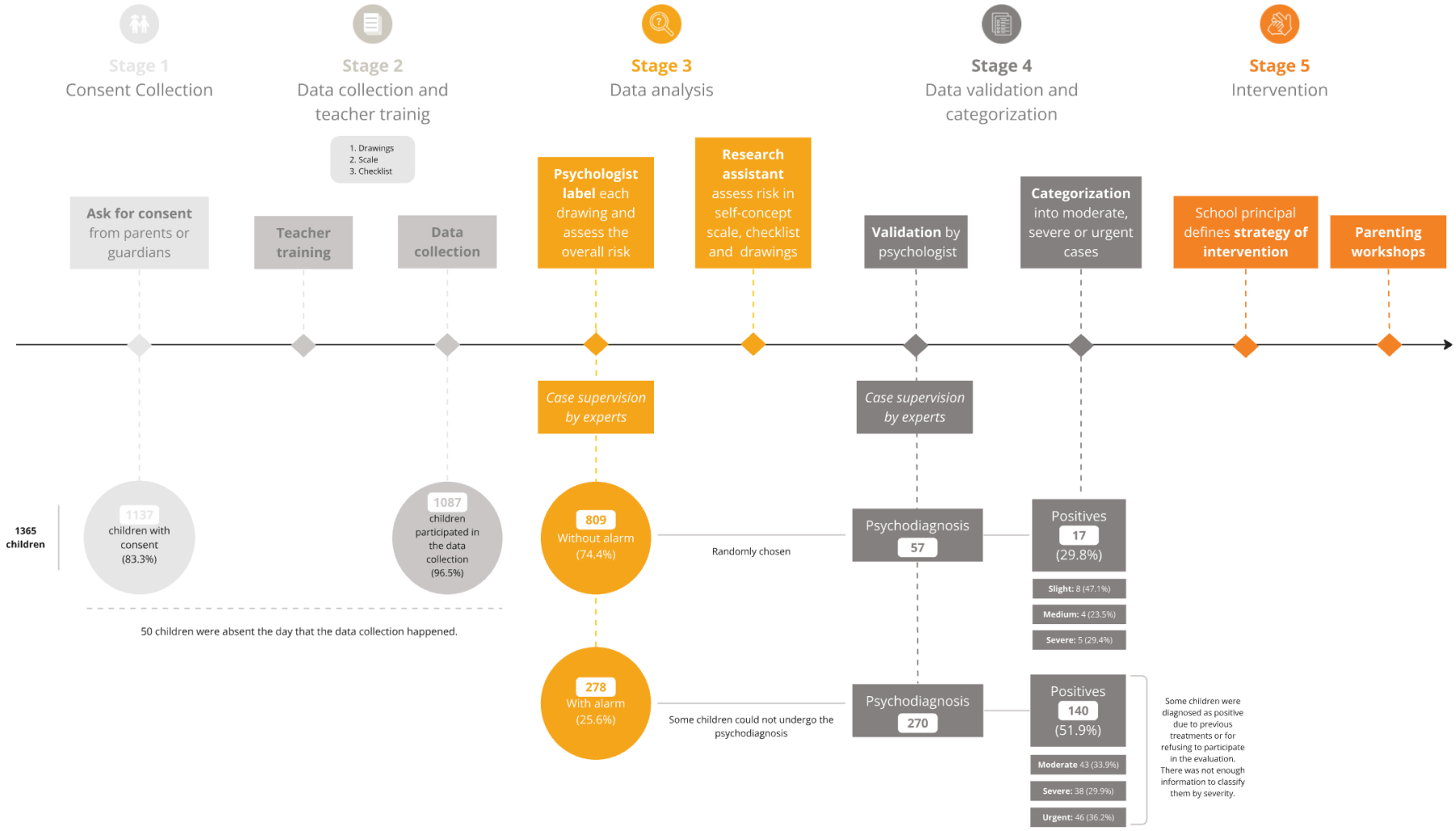
Stage 4: Application of Reference Test. In this stage, the comprehensive psychological assessments (as described in the previous section) were conducted. Prior to starting these assessments, psychologists participated in a virtual course provided by Fundación Salud Activa, focusing on report writing. Psychologists also received ongoing personalized and group support throughout this stage. Weekly group supervision sessions were held to discuss specific cases and address practical aspects of report preparation.

Cases were categorized as either positive or negative for maltreatment risk. Positive cases were further classified into three categories based on the severity of the situation and the frequency of violent incidents: urgent, severe, or moderate. Additionally, cases were categorized by the type of maltreatment experienced, including physical abuse, psychological abuse, neglect, sexual abuse, and exposure to violence. Based on these assessments, detailed feedback was provided to the schools, and individualized care plans were developed for each positive case.[19]

During the validation phase, of the 278 children identified with a red flag, 8 were unable to participate in the assessments: 4 were absent during the scheduled dates without the possibility of rescheduling, 2 had transferred to different schools after the diagnostic tools were applied but before assessments could be conducted, and 2 were excluded due to cognitive disabilities, which fell outside the protocol's scope. No attrition occurred in the randomly selected group of children without a red flag. No indeterminate results were reported for the reference standard. Cases in which the reference standard could not be completed were excluded from the diagnostic accuracy analysis; no imputation procedures were applied. No adverse events or distress requiring clinical response were reported during the administration of the comprehensive psychological assessments.

[19] Although not a core component of the diagnostic accuracy assessment, the final stage of implementation involved supporting schools in providing interventions for children identified as experiencing violence, in accordance with local protocols and regulations. The schools held primary responsibility for these interventions, while Fundación Abrazar provided support and training as needed to follow established procedures. This stage included monitoring and supporting all identified cases, with interventions tailored based on the severity and type of violence. For urgent cases, schools were responsible for actions such as reporting to child protection agencies, reducing contact with the abusive adult, or facilitating psychological treatment, with guidance from the research team when necessary. For severe and moderate cases, interventions included working with guardians, extending school hours, establishing support networks, and inviting families to participate in positive parenting workshops, with the research team offering additional support and training as required.

Figure 1: *Participant Flow and Implementation Stages of the Study*



Note: Authors' elaboration. The diagram shows the five implementation stages and the transitions between them. It also includes the number of participants retained at each step and the categories used for validation.

Importantly, the psychologists conducting these assessments were blinded to the responses from the three diagnostic tools and were unaware of whether the assessed children had been flagged as at risk.

The reference assessments in Stage 4 were conducted soon after the screening tools were applied in Stage 3 (up to 4 weeks), and interventions began only after the reference evaluations were completed. No clinical actions were taken between the index tests and the reference standard.

5 Empirical Design

To evaluate the diagnostic accuracy of the three diagnostic tools we assess their precision and sensitivity. We also use the results of the comprehensive psychological assessments to estimate the prevalence of maltreatment among our population.

Key variables and metrics. We define our main variables as follows:

Detection tool outcome. Let $X_{t,i}$ be the outcome of detection tool t on participant i. The variable takes the following values:

$$X_{t,i} = \begin{cases} 1 & \text{if the tool flags a potential case of maltreatment (red flag)} \\ 0 & \text{if the tool does not flag a potential case (no red flag)} \end{cases}$$

Comprehensive psychological assessment outcome. Let Y_i be the outcome of the comprehensive psychological assessment of participant i. The variable takes the following values:

$$Y_i = \begin{cases} 1 & \text{if maltreatment is found} \\ 0 & \text{if no maltreatment is found} \end{cases}$$

From these variables, we derive our primary metrics:

Precision, also known as Positive Predictive Value (PPV), is a central metric in our analysis. It is defined as the proportion of children correctly identified as experiencing maltreatment among all children flagged by the detection tool. Mathematically, this is expressed as $P(Y_i = 1|X_{t,i} = 1)$. Precision measures the accuracy of our tools among those flagged as at-risk individuals. A high precision value indicates that when our tool flags a child, there's a high likelihood that the child is indeed experiencing maltreatment, which is crucial for efficiently allocating limited intervention resources.

Sensitivity, also referred to as the True Positive Rate, is our second primary metric. It is defined as the proportion of children experiencing maltreatment who are correctly flagged by the detection tool. Mathematically, this is expressed as $P(X_{t,i} = 1|Y_i = 1)$. Sensitivity assesses our tools' ability to identify all potential cases of maltreatment. A high sensitivity indicates that the tool is effective at catching most cases of maltreatment, which is crucial for comprehensive child protection. However, it is important to note that increasing sensitivity often comes at the cost of reduced precision, a trade-off that must be

carefully considered in the context of limited available resources for follow-up and intervention.

Prevalence is another key metric, representing the proportion of children in the population experiencing maltreatment or neglect. It is expressed as $P(Y_i = 1)$, the probability that a randomly selected child from the population is experiencing maltreatment. Understanding prevalence is fundamental to grasping the scale of the problem. It provides context for interpreting other metrics and helps in assessing the overall impact of intervention programs.

Sampling approach. Our study employs a two-stage sampling approach designed to prioritize an accurate estimation of precision while still allowing for estimates of prevalence and sensitivity. Children were first screened using the diagnostic tools.[20] Selection for the comprehensive psychological assessment, which served as the reference test, was based on two criteria: first, all children who triggered a "red flag" in any of the diagnostic tools were selected for assessment, and second, a random subsample of children who did not trigger a red flag were also selected for assessment. To formalize our sampling strategy, we define the following variables:

Selection for the reference test. Let R_i indicate whether child i was selected for the comprehensive psychological assessment. The variable takes the following values:

$$R_i = \begin{cases} 1 & \text{if selected for comprehensive psychological assessment} \\ 0 & \text{if not selected for comprehensive psychological assessment} \end{cases}$$

Probability of selection for the reference test. The probability of selection for the reference test (π_i) is given by:

$$\pi_i = P(R_i = 1) = \begin{cases} 1 & \text{if } X_{t,i} = 1 \text{ for any } t \text{ (all red flag cases)} \\ \tilde{\pi} & \text{if } X_{t,i} = 0 \,\forall\, t \text{ (non-red flag cases)} \end{cases}$$

where $\tilde{\pi}$ represents the sampling fraction of non-red flag cases.

This design allows us to accurately estimate precision by assessing all red-flagged cases, while at the same time estimate prevalence and sensitivity through the inclusion of a random sample of non-flagged cases. Additionally, it provides estimates of maltreatment rates among non-flagged children, enabling crucial comparisons between flagged and unflagged groups and thereby validating the discriminatory power of our detection tools.

5.1 Sample size rationale

We conducted sample size calculations to ensure the statistical robustness of our findings. First, for the estimation of the tools' precision, we determined that assessing 267 children who triggered a red flag would provide a reliable estimate with a margin of error of less than 6 percentage points at a 95% confidence level. Then, to assess the tools' effectiveness in differentiating between at-risk and not-at-risk children, we calculated that assessing 58 children from the non-flagged group would provide sufficient

[20] As indicated in the previous section, the study population consists of students in the school who provided consent, which totaled 1137 students. Out of this sample, 50 students were absent on the day of the screening. The rest (96.5%) were screened by at least one of the tools.

statistical power to detect meaningful differences between these groups.

These calculations were based on standard statistical methods for estimating population proportions and comparing two independent proportions. Detailed calculations, including the specific formulas and assumptions used, are provided in Annex 5.

This sampling approach allows for the simultaneous pursuit of multiple objectives. First, it fulfills the ethical obligation to verify all potential risk cases identified by the diagnostic tools. Second, it enables the estimation of the tools' precision with sufficient statistical power. Third, it allows for comparisons between flagged and unflagged groups, thereby validating the discriminatory capacity of the instruments. Finally, it provides estimates of overall prevalence and sensitivity, albeit with some limitations in accuracy.

5.2 Estimation Procedures

Given the selective nature of the assessment process, weighted estimators are employed to calculate the key metrics: prevalence, precision, and sensitivity. The following notation is used throughout:

- m is the total number of children in the study.
- R_i is the selection indicator, equal to 1 if child i was selected for psychological assessment and 0 otherwise.
- π_i is the probability that child i was selected for assessment.
- X_i is the outcome of the diagnostic tool(s) for child i (1 if flagged, 0 otherwise).
- Y_i is the outcome of the comprehensive psychological assessment for child i (1 if maltreatment is found, 0 otherwise).

1. **Prevalence Estimation:** The prevalence of maltreatment, denoted by $\lambda = P(Y = 1)$, represents the proportion of children in the population who experienced abuse or neglect. Given the stratified nature of the sample, it is estimated using a Horvitz–Thompson-type estimator that adjusts for the sampling probabilities:

$$\lambda = \mathbb{E}[Y] = \mathbb{E}\left[\frac{R}{\pi} \cdot Y\right] \quad \text{estimated by} \quad \hat{\lambda} = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{R_i}{\pi_i} \cdot Y_i\right)$$

   This estimator weights each observed case by the inverse of its selection probability, allowing inference on the full study population.

2. **Precision Estimation:** Precision, or Positive Predictive Value, is defined as the proportion of children who were correctly identified as experiencing maltreatment among those flagged by at least one detection tool:

$$p_1 = P(Y = 1 \mid X = 1) \quad \text{estimated by} \quad \hat{p}_1 = \frac{\sum_{i=1}^{m}(R_i \cdot X_i \cdot Y_i)}{\sum_{i=1}^{m}(R_i \cdot X_i)}$$

Since this metric is computed within the subgroup of children who were both flagged and assessed, it does not require weighting by selection probabilities.

3. **Sensitivity Estimation:** Sensitivity, or True Positive Rate, represents the proportion of actual maltreatment cases that were successfully flagged by the tools. It is expressed as:

$$\theta = P(X = 1 \mid Y = 1) = \frac{\mathbb{E}\left[\frac{R}{\pi} \cdot X \cdot Y\right]}{\mathbb{E}\left[\frac{R}{\pi} \cdot Y\right]} \quad \text{estimated by} \quad \hat{\theta} = \frac{\sum_{i=1}^{m}\left(\frac{R_i}{\pi_i} \cdot X_i \cdot Y_i\right)}{\sum_{i=1}^{m}\left(\frac{R_i}{\pi_i} \cdot Y_i\right)}$$

This ratio adjusts for the sample design and accounts for the probability of selection, providing an unbiased estimate of tool sensitivity.

For each of these metrics, we derive standard errors and construct confidence intervals using the delta method. Detailed mathematical derivations for these calculations can be found in Annex 6.

In the following sections, we present the results of our analysis, including estimates of prevalence, precision, and sensitivity for various configurations of our diagnostic tools.

6 Results: Prevalence, Precision and Sensitivity

This section presents the results of our empirical analysis for prevalence, precision and sensitivity of these tools and compares their detection capabilities to those of the current status quo practices.[21]

6.1 Prevalence estimates

The comprehensive psychological assessments reveal that 35.3% (95% CI [25.9, 44.7]) of the sample experienced some form of maltreatment or neglect, illustrating the pervasive nature of this issue, which is consistent with other studies (WHO, 2024). In this sample, 14.5% (95% CI [7.7, 21.3]) of participants experienced moderate cases of sporadic violence, while 8.8% (95% CI [3.7, 13.9]) encountered severe, recurrent violence occurring at least twice a year, and 10.8% (95% CI [5.2, 16.4]) faced urgent situations with potential risks of death or serious harm.

In terms of abuse types, neglect and physical abuse were the most prevalent, estimated at 17.0% (95% CI [10.1, 23.9]) and 16.6% (95% CI [9.7, 23.5]) of the total sample, respectively. Emotional abuse was estimated to affect 15.0% (95% CI [8.4, 21.6]) of the participants. Gender-based violence was estimated at 7.4% (95% CI [2.4, 12.4]) of the sample. Sexual abuse, including both indirect and direct forms, was less prevalent, as detailed in Figure 2.[22]

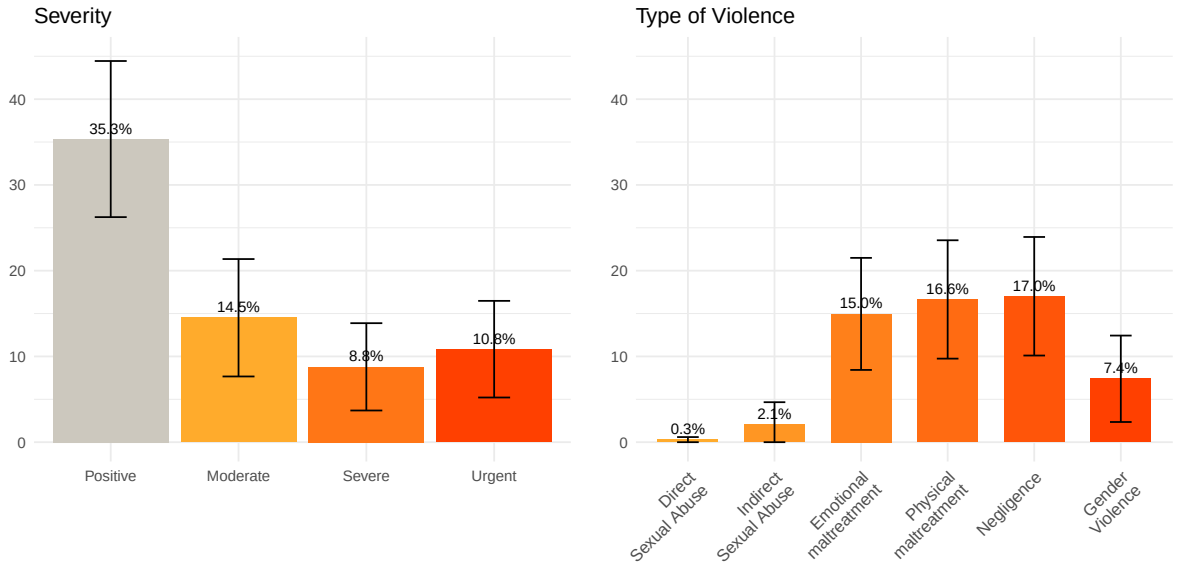
These figures represent estimates based on the comprehensive psychological assessments used as the reference test, which may be conservative for some types of abuse. Particularly for sexual abuse, actual

[21] The thresholds used to classify red flags for each tool were pre-specified as defined in Section 3. We also conducted exploratory analyses of diagnostic accuracy across alternative thresholds for each tool, as shown in Annex 7. These analyses were not pre-specified.

[22] Note that totals do not add to 35.3% because different types of maltreatment might co-occur.

figures might be higher due to the difficulties of detecting this type of abuse in the limited timeframe of the assessments. Despite this factor, the identified prevalence rates provide a stark indication of the extent of maltreatment and neglect in our population.

Figure 2: *Estimates of the Percentage of Maltreatment Cases, by Severity and Type of Maltreatment*



Source: Authors' calculations. **Note:** Maltreatment cases are identified through comprehensive psychological assessments (including prior school identification). Bars represent point estimates; whiskers indicate 95% confidence intervals.

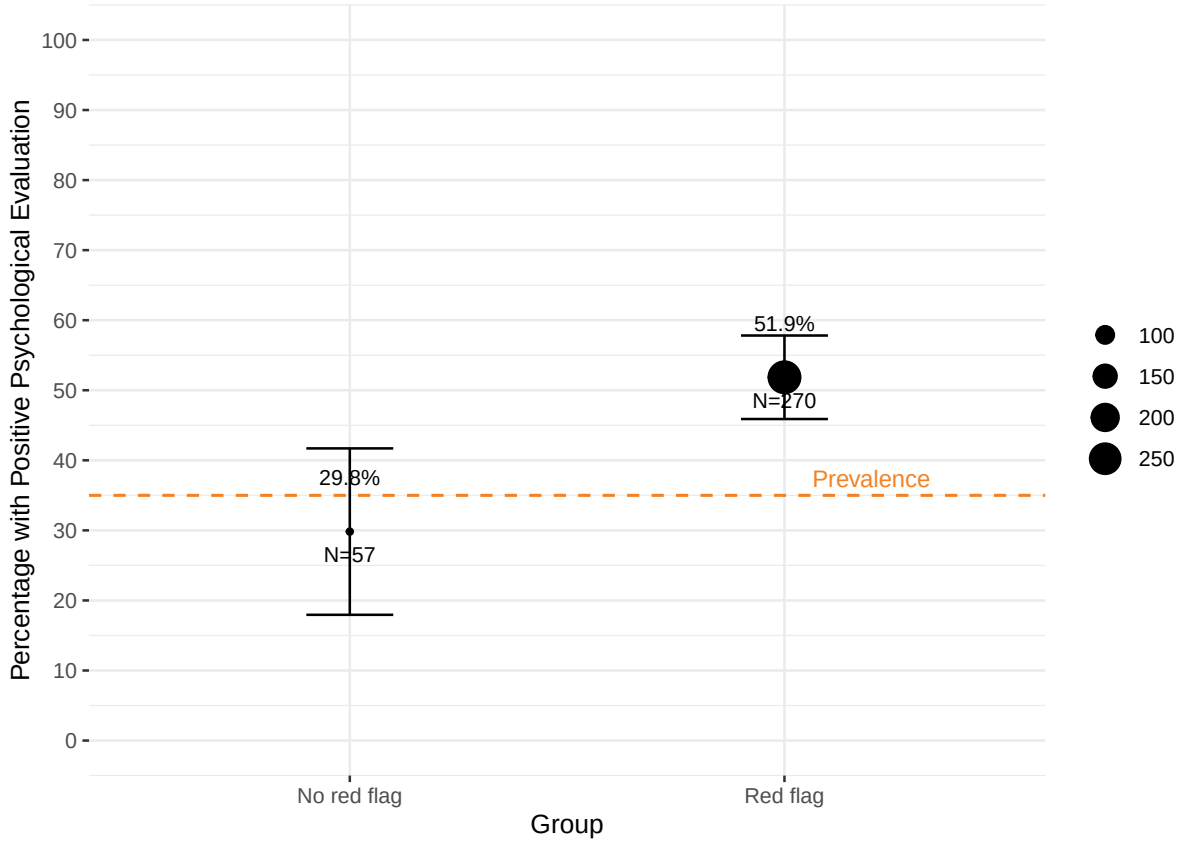
6.2 Precision

Out of the 1087 children and adolescents in our sample, 278 (25.6%) were red flagged in at least one of the three diagnostic tools and 270 of them underwent a subsequent psychological assessment. Within this group of children and adolescents flagged by the diagnostic tools, 51.9% (equating to 140 cases) were assessed to have experienced maltreatment or neglect by the psychological evaluations used as reference test. Conversely, in the random group of 57 children and adolescents without an alert who were assessed, 29.8% (representing 17 cases) were identified as having been subjected to maltreatment or neglect. The odds ratio for detecting maltreatment in red flagged cases versus no red flagged cases was 1.74, with a 95% confidence interval ranging from 1.45 to 2.02 (see Figure 6). The difference in the rates of maltreatment or neglect between the children with a red flag and those without a red flag in the diagnostic tools is statistically significant (at the 5% significance level). This indicates that the tools can jointly differentiate between children at risk of abuse or mistreatment from those who are not. Moreover, this significant difference is maintained when comparing the prevalence in the group with a red flag against the estimated overall prevalence rate of 35.3%. In other words, the tools work better in identifying children at-risk of maltreatment or neglect than a protocol consisting of a random sampling of the population. Figure 3 displays these results and their 95% confidence interval.

In the contexts of limited resources in which protection systems and school usually operate, it is necessary to prioritize attention or even the administration of comprehensive diagnostic tests for some children. The set of diagnostic tools can serve to enhance the correct allocation of such resources. However, they

are not infallible, as shown by the presence of false positives. These results indicate that while the diagnostic tools can be beneficial for targeting children to receive further attention or comprehensive assessment, they should not be used as definitive diagnostic instruments. These tools could serve as a preliminary step in a multifaceted approach to child protection, which, consistent with best practices, should be applied in a way not to stigmatize families or children.

Figure 3: *Percentage of Children With Confirmed Maltreatment in Groups With and Without a Red Flag in the Diagnostic Tools, and Overall Prevalence*



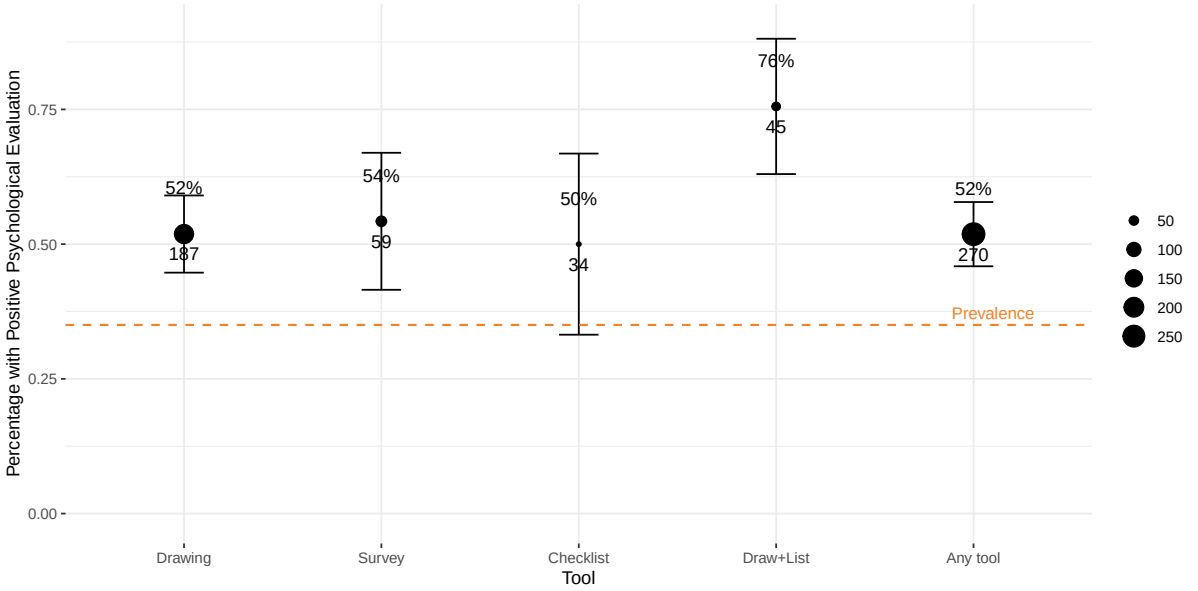
Source: Authors' calculations. **Note:** Maltreatment cases are determined through comprehensive psychological assessments (including prior school identification). The dot shows the percentage of cases with confirmed maltreatment. Dot size represents the sample size used to calculate prevalence among children with and without a red flag. Whiskers indicate 95% confidence intervals.

We further study the precision for each tool separately and for different combinations of the tools. We explore combinations of tools considering when red flags are simultaneously triggered by multiple tools ('and' conditions) as well when a red flag is triggered by any tool ('or' condition). Details on all combinations are available in Annex 1. In what follows, we will refer to the projective drawings as 'Drawings,' the Piers-Harris Self-Concept Scale as 'Survey,' and the teacher's checklist as 'Checklist.'

Figure 4 reports the precision for each tool and for one combination (Draw + List) chosen for its high precision. The graph displays the precision metric with its 95% confidence interval. The dot size represents the number of cases flagged by each tool. For example, among all children, 187 were flagged for signs of violence or abuse in their Drawings, with 51.9% (95% CI [44.7, 59]) of these cases assessed through psychological evaluations as true positives.

Figure 4 shows that each diagnostic tool achieves a precision rate of at least 50%. This indicates that

Figure 4: *Precision of the Different Diagnostic Tools*



Source: Authors' calculations. **Note:** Maltreatment cases are determined through comprehensive psychological assessments (including prior school identification). The dot position shows the percentage of cases with confirmed maltreatment. Dot size represents the number of cases assessed with a red flag for each tool or combination of tools. Whiskers indicate 95% confidence intervals.

for each tool, at least half of the red-flagged cases were assessed as positive by the comprehensive psychological evaluations used as the reference test. The precision rates of all tools exceed the overall estimated prevalence rate of 35.3% (95% CI [25.9, 44.7]), with only the precision of Drawings showing a statistically significant difference from this rate. This implies the following odds ratios when comparing each tool to the overall prevalence: 2.0 (95% CI [1.40, 2.88]) for Drawings, 2.18 (95% CI [1.25, 3.80]) for the Survey, and 1.86 (95% CI [0.91, 3.78]) for the Checklist.[23]

While the precision of each individual tool is close to 50%, the combined use of Projective Drawings and the Teacher's Checklist achieves a precision rate of 75.6%. This indicates that over three-quarters of children flagged by both Projective Drawings and Teacher's Checklist were assessed as positive cases by the psychological evaluations used as reference tests.

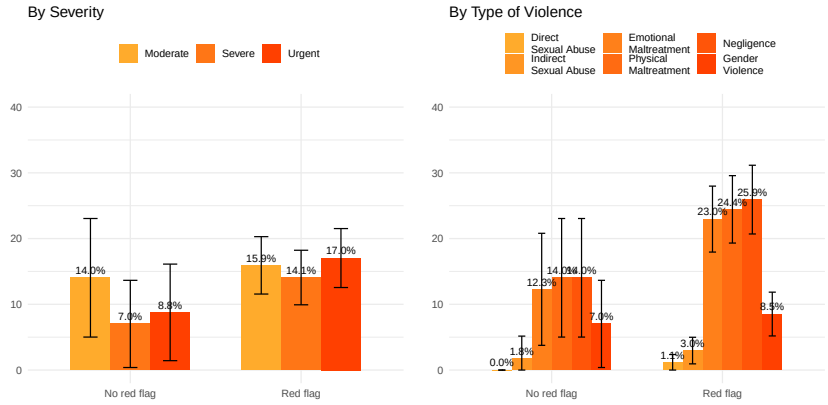
Precision across different severities of maltreatment. Figure 5 (left panel) shows the prevalence of different severities of maltreatment among children with and without a red flag in the diagnostic tools. The data reveals that the different severities of maltreatment are more prevalent among children with a red flag, with greater differences observed in cases of severe and urgent maltreatment. Specifically, among children not flagged by a red flag, 14% were assessed to have experienced moderate maltreatment, 7% severe, and 8.8% urgent maltreatment. In contrast, among children with a red flag, these percentages increase to 15.9% for moderate, 14.1% for severe, and 17.0% for urgent cases, as assessed by the psy-

[23]The number of Surveys and Checklist with a red flag is low in comparison to that of Drawings because in both cases, not all children are considered. Surveys were only considered for children in 5th grade and above, a total of 370 children completed the survey, out of these, 60 resulted with a red flag based on their responses. All children with a red flag underwent the psychological evaluation, though in the case of the red flagged cases based on the Survey, one child was absent the days the psychological evaluation took place and is therefore not considered in the sample. This results in the 59 cases reported in this analysis. Regarding the Checklist, we are only considering the cases in which this tool was applied to all kids regardless of having a red flag in previous tools (Drawings or Survey) which happened in the second round of the intervention. There were 331 children in this round, and out of these, 34 resulted in a red flag in the Checklist. These are the ones plotted in the graph.

chological evaluations. These differences mean that the odds of detecting severe or urgent maltreatment are significantly higher in children with a red flag compared to those without, with an odds ratio of 2.4 (95% CI [1.1, 3.7]). In summary, it is approximately twice as likely to find cases of maltreatment among children flagged by any of the three diagnostic tools compared to those not flagged (see Figure 6).

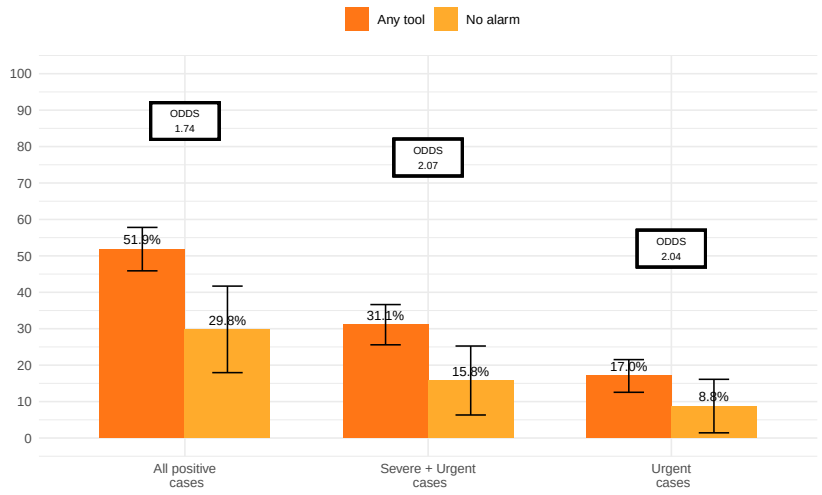
We further evaluated the precision of the combination for different severities of maltreatment of projective drawings and Teacher's Checklist, which was identified as the most precise combination. For children flagged by both tools, the percentages of children assessed as experiencing different severities of maltreatment were: 13% moderate, 23% severe, and 40% urgent. Conversely, for children flagged by at most one of these tools, the respective percentages were 14% for moderate, 7% for severe, and 9% for urgent maltreatment. In other words, the odds ratio for detecting severe and urgent cases when both tools trigger a red flag, compared to when at most one tool triggers a red flag, is 3.98, with a 95% confidence interval of 3.76 to 4.18. For the most urgent cases, the odds ratio increases to 4.51, with a confidence interval of 4.08 to 4.93 (see Figure 7).

Figure 5: *Maltreatment Prevalence Among Children With and Without a Red Flag in at Least One Tool, by Severity and Type of Maltreatment*



Source: Authors' calculations. **Note:** Maltreatment cases are determined through comprehensive psychological assessments (including prior school identification). Bars represent the percentage of cases with confirmed maltreatment; whiskers indicate 95% confidence intervals.

Figure 6: *Maltreatment Prevalence Among Children With a Red Flag in at Least One Tool and Without a Red Flag in Any Tool, by Severity and Type of Violence*



Source: Authors' calculations. **Note:** Maltreatment cases are determined through comprehensive psychological assessments (including prior school identification). Bars represent the percentage of cases with confirmed maltreatment; whiskers indicate 95% confidence intervals.

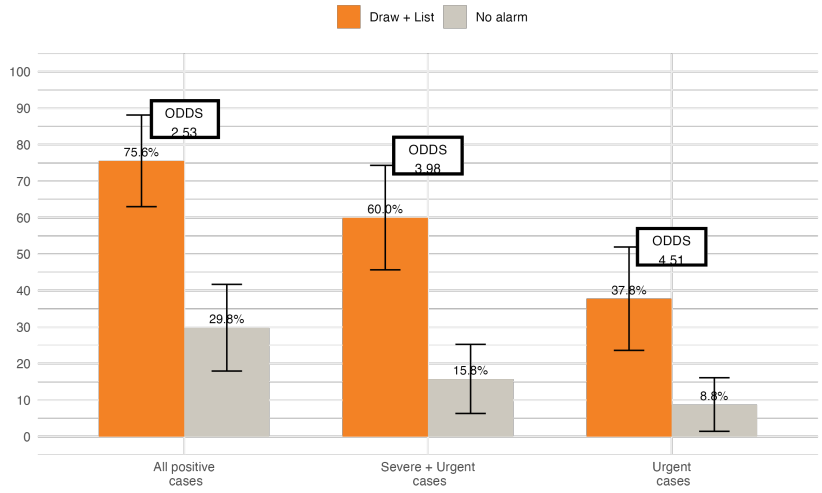
Precision across different types of maltreatment. Figure 5 (right panel) illustrates the prevalence of different types of maltreatment among children with and without a red flag in the diagnostic tools. While prevalence rates have wide confidence intervals and differences across groups are not statistically significant, findings indicate that all assessed types of maltreatment are more prevalent among children with a red flag than those without. This result is consistent with the diagnostic tools being able to identify cases of different types of maltreatment. Specifically, among children not flagged by a red flag, the rates of maltreatment were found to be 1.8% for sexual abuse, 12.3% for emotional maltreatment, 14% for physical maltreatment, 14% for negligence, and 7% for gender violence. Conversely, among children with a red flag, the prevalence increases to 4.1% for sexual abuse, 23% for emotional maltreatment, 24.4% for physical maltreatment, 25.9% for negligence, and 8.5% for gender violence, as assessed by the psychological evaluations.

Precision across different subpopulations. Figure 8 shows the rates of maltreatment as assessed by the psychological assessments for both the red flagged and no red flagged groups, segmented by sex, age, and socioeconomic status (SES) (Figures A.3 and A.4 in Annex 7 present this information disaggregated by severity and type of abuse). While the small sample size limits our ability to identify statistically significant differences, some patterns emerge.

Regarding sex (Figure 8, left panel), boys show a higher rate of positive psychodiagnostics in both groups. Specifically, 25.8% of girls and 34.6% of boys in the group without a red flag were assessed as experiencing maltreatment by the comprehensive psychological evaluations. These rates increase to 48.7% for girls and 54.1% for boys in the group with a red flag. On the one hand, these results suggest higher incidence of violence among boys. On the other hand, the increased prevalence in both boys and girls within the red flagged groups is consistent with the diagnostic tools' being able to effectively identify risks across different sexes.

Regarding age patterns (Figure 8, center panel), we observe a decreasing trend in the prevalence of

Figure 7: *Maltreatment Prevalence Among Children With a Red Flag and Without a Red Flag in Both Projective Drawings and Teacher's Checklist, by Severity*



Source: Authors' calculations. **Note:** Maltreatment cases are determined through comprehensive psychological assessments (including prior school identification). Bars represent the percentage of cases with confirmed maltreatment; whiskers indicate 95% confidence intervals.

maltreatment as age increases. While not statistically significant, this trend could imply that younger children are either more exposed to abuse or that the psychological assessments are more attuned to identifying signs of abuse in this age group.[24] In terms of precision, although not statistically significant, there is a consistently higher prevalence of identified maltreatment among the children flagged with a red flag across all age groups (relative to those without a red flag). This observation supports the effectiveness of the diagnostic tools in recognizing signs of abuse irrespective of age.

Lastly, in terms of SES (Figure 8, right panel), our analysis does not reveal a clear pattern of maltreatment prevalence across different socioeconomic groups. Nonetheless, the data indicate a higher prevalence of maltreatment among children with a red flag in all socioeconomic groups, suggesting that the diagnostic tools are effectively identifying risks across diverse SES levels.

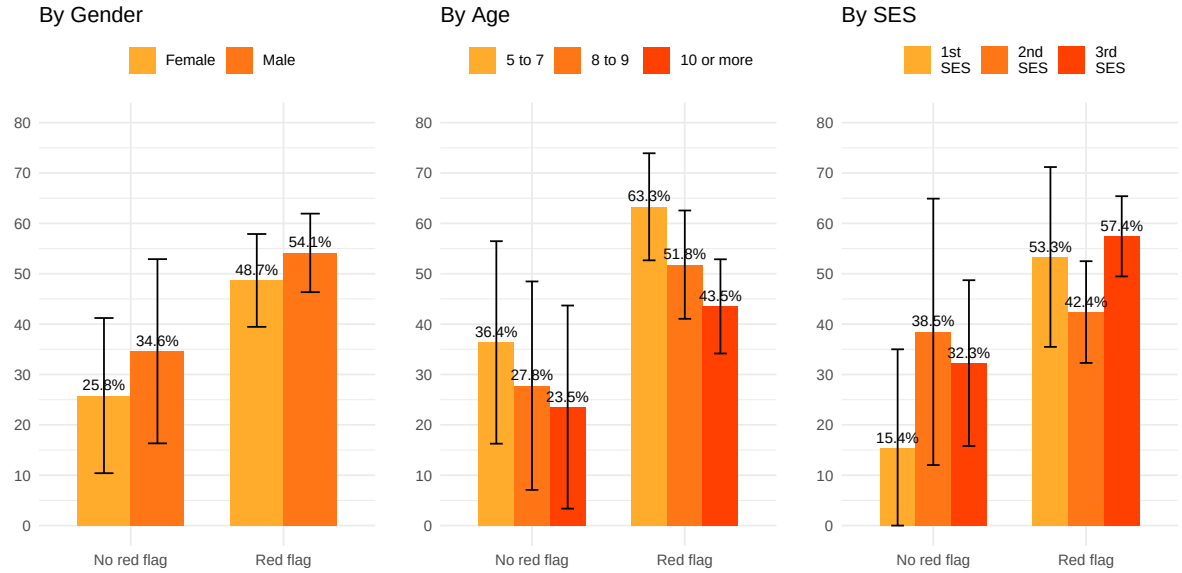
6.3 Sensitivity and Comparison against the Status Quo

Beyond precision, it is essential to analyze the sensitivity of the diagnostic tools used in our study. Sensitivity measures the ability of a tool or combination of tools to identify positive cases, providing an essential metric for evaluating their effectiveness in uncovering the extent of maltreatment. Typically, there is a trade-off between precision and sensitivity, where focusing on achieving higher precision can sometimes lead to missing some positive cases, reducing sensitivity.

Figure 6 presents both precision and sensitivity for the three individual tools and the combination of Projective Drawings and Teacher's Checklist, selected for its high precision. Utilizing all three tools

[24] Drawing on the clinical experience of two psychologists on the research team—each with over 30 years of professional practice—it is observed that younger children in primary school are more likely to disclose experiences of abuse. This tendency may be attributed to their limited awareness of social stigma and their reduced sense of shame compared to older children or adults. Younger children may not fully understand the nature of the abuse or its implications, and as a result, they may disclose such experiences more openly, sometimes describing them in the context of a game.

Figure 8: *Maltreatment Prevalence Among Children With a Red Flag in at Least One Tool and Without a Red Flag in Any Tool, by Gender, Age, and Socioeconomic Status*



Source: Authors' calculations. **Note:** Maltreatment cases are determined through comprehensive psychological assessments (including prior school identification). Bars represent the percentage of cases with confirmed maltreatment; whiskers indicate 95% confidence intervals.

collectively resulted in the highest sensitivity, identifying 36.8% of positive cases. When assessed individually, the sensitivity rates varied significantly: the Teacher's Checklist identified 13% of the cases, the Projective Drawings identified 25.5%, and the surveys identified 30%.

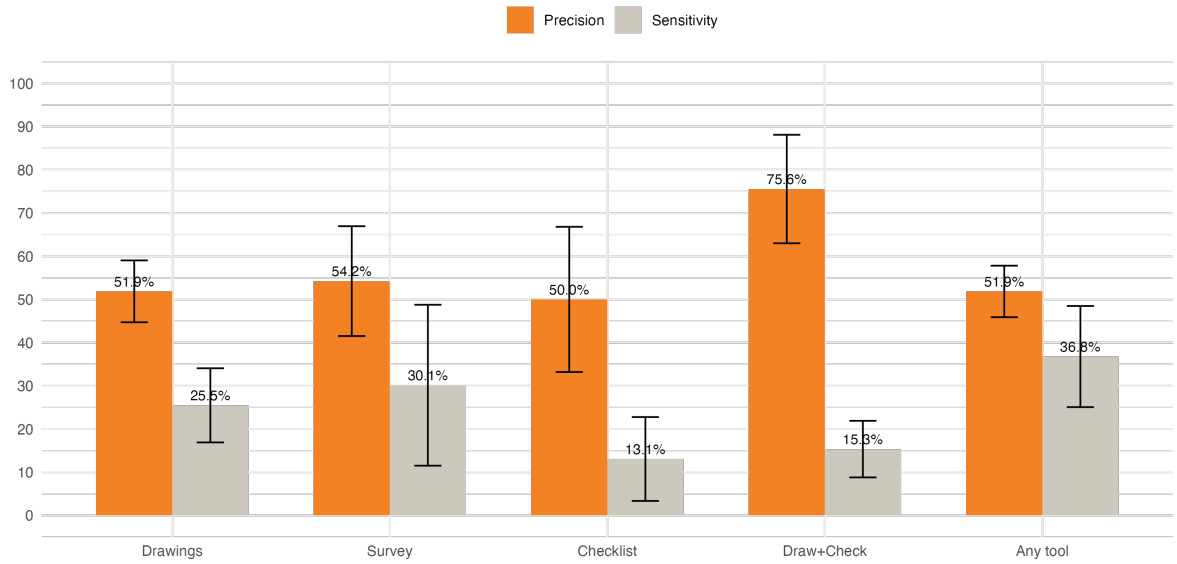
When assessing the effectiveness of the set of diagnostic tools across different severities of maltreatment, the sensitivity rates demonstrate a better performance in identifying more intense cases. Specifically, the sensitivity for severe and urgent maltreatment combined is 39.7%, and for urgent maltreatment alone, is 39.4%[25]. These rates suggest that approximately 40% of children experiencing the most severe forms of maltreatment are successfully identified by these tools. However, as mentioned above, they also identify some individuals who are not experiencing maltreatment.

In the previous section, we emphasized the high precision achieved by combining Projective Drawings and the Teacher's Checklist. However, it is important to note that this combination also exhibits a relatively low sensitivity rate, identifying only 9% of all positive cases. This observation is detailed further in Table A.1 in Annex 7, which includes the full range of instruments and their combinations. The trade-off illustrated here is significant: while this combination reduces the number of false positives, it also results in a higher number of false negatives, underscoring the inherent challenges in balancing these two aspects of diagnostic accuracy.

Precision and sensitivity are critical for evaluating the effectiveness of diagnostic tools, yet it is also important to understand how these tools may bolster ongoing efforts within schools to identify and support children experiencing maltreatment. In two of the participating schools, we compared the performance of our tools against the existing detection and intervention processes, collecting data on

[25]Refer to Annex 7, Tables A.2 trough A.7 to see how sensitivity changes for different severities and types of abuse detected (negligence, physical and emotional abuse) and for each tool and combination of diagnostic tools.

Figure 9: *Precision and Sensitivity of the Detection Tools*



Source: Authors' calculations. **Note:** Precision is defined as the percentage of red-flagged cases confirmed as maltreatment through comprehensive psychological assessments (or prior school identification). Sensitivity represents the percentage of confirmed maltreatment cases detected by the tool. Bars show point estimates; whiskers indicate 95% confidence intervals.

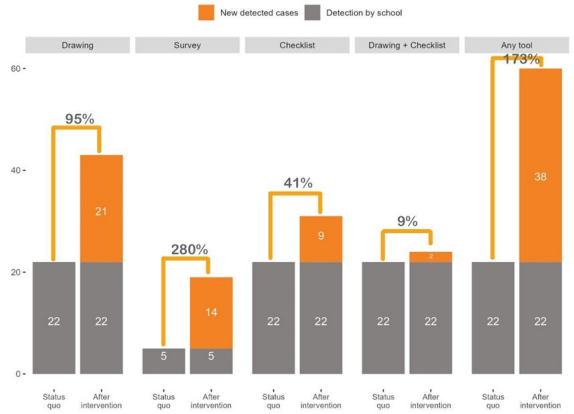
treatment or intervention status at the time of the study.[26] Figure 7 presents the number of children identified by the schools and those detected by our tools, highlighting the proportion of new cases the diagnostic tools uncovered.

The implementation of the trio of diagnostic tools resulted in a 173% increase in the number of identified cases, adding 38 cases overall. Breaking this down by individual tools, Projective Drawings uncovered 21 cases, representing a 95% increase. The self-concept scales, administered to older children, led to 14 additional cases, marking a 280% increase compared to the five cases initially identified among this population. The Teacher's Checklist contributed to detecting 9 cases (a 41% increase). The combined use of projective drawings and the Teacher's Checklist revealed 2 new cases, a 9% increase relative to the existing detection rates.

In short, while the tools' sensitivity rates indicate that they capture less than half of the universe of maltreatment cases, they allowed for an improvement over the current detection standards in the context of our study, uncovering new cases beyond the schools' initial identification.

[26] In the first round (three schools in the City of Buenos Aires), information about previously confirmed cases was gathered through consultation with the school principal at the point when children were selected for psychological assessment. Therefore, we only have information about cases among those who were flagged by the diagnostic tools or randomly selected to do the comprehensive psychological assessment. In contrast, during rounds two and three (in the Córdoba schools), a more systematic approach was adopted: a complete list of children with confirmed cases of maltreatment or ongoing treatment was requested from school staff prior to the application of the diagnostic tools. In the analysis presented in the following paragraphs, we focus only on the two schools in Cordoba as we have the complete list of cases identified by the schools.

Figure 10: *New Detected Maltreatment Cases With a Red Flag in Different Diagnostic Tools*



Source: Authors' calculations. **Note:** New cases refer to those identified through the diagnostic tools that had not been previously detected by the school. Bars show counts of cases previously identified by the school (gray) and new detections by each tool (orange). Percentages indicate the increase in detected cases relative to prior identification.

7 Conclusions

The World Health Organization (WHO) estimates that 23% of children and adolescents experience physical abuse in their lifetime, with sexual abuse affecting 18% of girls and 8% of boys (WHO, 2024). In Latin America and the Caribbean, recent estimates indicate that two out of three children under the age of fifteen experience some form of violence as a disciplinary method, and half suffer physical punishments at home (UNICEF, 2022). Such maltreatment not only infringes on children's rights but also impairs their well-being and hampers their development. The high prevalence and profound impacts of child maltreatment make it a severe public policy challenge.

Effective public policy is limited by a pervasive culture of silence and institutional constraints, which contribute to a substantial underreporting of maltreatment cases. This study addresses the critical issue of underreporting in child maltreatment by evaluating the diagnostic accuracy of scalable screening tools within educational settings—namely, Projective Drawings, Piers-Harris Self-Concept Scales, and Teacher's Checklist. Our findings indicate that these trio of tools can enhance the identification of maltreatment cases. Their precision rate is over 50%, indicating that more than half of the red flagged cases corresponded with cases of maltreatment assessed by subsequent psychological evaluations. It is approximately two times more likely to find a case of maltreatment among those children with a red flag in at least one of the tools compared to those without a red flag. Furthermore, considering cases with red flags in both projective drawings and teacher's checklists showed a remarkable increase in precision, with 76% of true positive cases, thereby reducing the rate of false alarms. This way, the tools offer relevant informational value, particularly in resource-limited settings where prioritizing actions is crucial.

However, while informative, they also produce false positives and negatives. False negatives imply that the tools do not capture the full extent of the issue. Overall, the trio of tool have a sensitivity rate of about 36% to 40%, for more severe cases. While most cases are not detected, the tools nevertheless led to substantial increases in detection rates compared to existing school processes. Specifically, the implementation of these tools more than doubled the detection of maltreatment cases in participating

schools, highlighting their potential to significantly improve current practices.

Regarding false positives, our results indicate that the diagnostic tools should not be used as definitive diagnostic instruments. These tools could serve as a preliminary step in a multifaceted approach to child protection, which, consistent with best practices, should be applied in a way not to stigmatize families or children.

Our findings align with concerns raised by McTavish et al., 2020 regarding the challenges of relying on existing standardized tools for child maltreatment detection. As they observed, most tools evaluated to date have demonstrated limited diagnostic accuracy and are often inappropriate for broad application outside high-risk or clinical settings. By focusing on the school context and integrating tools that require minimal specialized training, our study seeks to offer a complementary approach that addresses some of the limitations identified in that systematic review, while acknowledging that these early detection tools are not substitutes for comprehensive psychological evaluation.

Overall, our study brings evidence that can help break the cyclical nature of underreporting and the systemic challenges that perpetuate it. The prevailing underestimation of child maltreatment leads to insufficient investment in protective systems, leaving them overwhelmed and less capable of managing existing cases or implementing effective detection and prevention strategies. To break this cycle, our findings advocate for the adoption of these early detection tools as part of a multifaceted approach to child protection. This approach should not only aim at enhancing detection but also at fostering a community-engaged, preventive framework that respects and protects without stigmatizing families or children.

Our study sought to address key methodological limitations noted in prior systematic reviews (Bailhache et al., 2013; McTavish et al., 2020). Specifically, many earlier studies applied the reference standard only to children who screened positive, introducing verification bias and limiting accurate estimates of false negatives. The study's design mitigated this by applying comprehensive psychological assessments not only to children flagged by the detection tools but also to a random sample of unflagged children. This approach allowed to estimate false negatives and provided a more robust assessment of diagnostic accuracy. Furthermore, the study included a broad age range within primary school settings and maintained strict separation between index tools and reference tests to minimize bias and enhance the validity of our findings.

Moving forward, it is imperative to continue rigorously testing the diagnostic accuracy of these tools in a variety of settings and explore ways to facilitate their at-scale implementation. Continuing to refine these tools and strategies will be critical to ensuring they can effectively contribute to the global effort to protect children and support their well-being in diverse educational and community environments.

References

Aberastury, A. (1962). *Teoría y técnica del psicoanálisis de niños* (Vol. 21). Paidós.

Algoritmos Públicos. (2021). Proyecto sistema de alerta temprana niñez. https://algoritmospublicos.cl/proyecto-sistema-alerta-temprana-ninez

Allen, B., & Tussey, C. (2012). Can projective drawings detect if a child experienced sexual or physical abuse? a systematic review of the controlled research. *Trauma, violence, & abuse*, *13*(2), 97–111.

Altemeier, W. A., O'Connor, S., Vietze, P., Sandler, H., & Sherrod, K. (1984). Prediction of child abuse: A prospective study of feasibility. *Child Abuse & Neglect*, *8*(4), 393–400. http://refhub.elsevier.com/S0145-2134(17)30352-6/sbref0010

American Professional Society on the Abuse of Children (APSAC). (2022). Apsac practice guidelines: Forensic mental health evaluations when child maltreatment is at issue (updated 2022) [Practice guideline].

American Psychological Association. (2024). Guidelines for psychological evaluations in child protection matters.

Assink, M., van der Put, C. E., Oort, F. J., & Stams, G. J. J. M. (2015). The development and validation of the youth actuarial care needs assessment tool for non-offenders (Y-ACNAT-NO). *BMC Psychiatry*, *15*, 36. http://refhub.elsevier.com/S0145-2134(17)30352-6/sbref0035

Averbuj, G. (2010). Maltrato infantil: Orientaciones para actuar desde la escuela. *Ministerio de Educación de la Nación*. http://www.edusalta.gov.ar/index.php/docman/secretaria-de-planeamiento-educativo/buenas-practicas-de-convivencia-institucional/2816-maltrato-infantil-orientaciones-para-actuar-desde-la-escuela/file

Ayling, N. J., Walsh, K., & Williams, K. E. (2020). Factors influencing early childhood education and care educators' reporting of child abuse and neglect. *Australasian Journal of Early Childhood*, *45*(1), 95–108.

Ayoub, C. C., & Milner, J. S. (1985). Failure to thrive: Parental indicators, types, and outcomes. *Child Abuse & Neglect*, *9*(4), 491–499. http://refhub.elsevier.com/S0145-2134(17)30352-6/sbref0045

Azzopardi, C., Eirich, R., Rash, C. L., MacDonald, S., & Madigan, S. (2019). A meta-analysis of the prevalence of child sexual abuse disclosure in forensic interviews. *Child Abuse & Neglect*, *93*, 291–304. https://doi.org/10.1016/j.chiabu.2019.05.023

Bailhache, M., Leroy, V., Pillet, P., & Salmi, L.-R. (2013). Is early detection of abused children possible?: A systematic review of the diagnostic accuracy of the identification of abused children. *BMC Pediatrics*, *13*(202). https://doi.org/10.1186/1471-2431-13-202

Baird, C., & Wagner, D. (2000). The relative validity of actuarial-and consensus-based risk assessment systems. *Children and Youth Services Review*, *22*(11), 839–871. http://refhub.elsevier.com/S0145-2134(17)30352-6/sbref0050

Ballús, E., Comelles, M. C., Pasto, M. T., & Benedico, P. (2023). Children's drawings as a projective tool to explore and prevent experiences of mistreatment and/or sexual abuse. *Frontiers in Psychology*, *14*, 1002864.

Barber, J. G., Shlonsky, A., Black, T., Goodman, D., & Trocmé, N. (2008). Reliability and predictive validity of a consensus-based risk assessment tool. *Journal of Public Child Welfare*, *2*(2), 173–195. http://refhub.elsevier.com/S0145-2134(17)30352-6/sbref0055

Benia, L. R., Hauck-Filho, N., Dillenburger, K., & Rossato, C. (2015). The nichd investigative interview protocol: A meta-analytic review. *Trauma, Violence, & Abuse*. https://doi.org/10.1177/1524838015584369

Benson, C., Fitzpatrick, M., & Bondurant, S. (2022). Beyond reading, writing, and arithmetic: The role of teachers and schools in reporting child maltreatment. *Journal of Human Resources*. https://doi.org/10.3368/jhr.0319-10084R2

Bianco, M., Watcher, P., Chiapparrone, N., & Müller, B. (2016). Abuso sexual en la infancia: Guía para orientación y recursos disponibles en CABA y Buenos Aires. http://feim.org.ar/wp-content/uploads/2016/08/Guia-ASI-2016.pdf

Blain, G. H., Bergner, R. M., Lewis, M. L., & Goldstein, M. A. (1981). The use of objectively scorable house-tree-person indicators to establish child abuse. *Journal of Clinical Psychology*, *37*(3), 667–673.

Bleger, J. (1964a). *La entrevista psicológica. su empleo en el diagnóstico y la investigación*. Universidad de Buenos Aires, Facultad de Filosofía y Letras, Departamento de Psicología.

Bleger, J. (1964b). *Temas de psicología: Entrevista y grupos*. Ed. Gedisa.

Brayden, R. M., Altemeier, W. A., Dietrich, M. S., Tucker, D. D., Christensen, M. J., McLaughlin, F. J., & Sherrod, K. B. (1993). A prospective study of secondary prevention of child maltreatment. *The Journal of Pediatrics*, *122*(4), 511–516.

Bright, M. A., Roehrkasse, A., Masten, S., Nauman, A., & Finkelhor, D. (2022). Child abuse prevention education policies increase reports of child sexual abuse. *Child Abuse & Neglect*, *134*, 105932. https://doi.org/10.1016/j.chiabu.2022.105932

Camasso, M. J., & Jagannathan, R. (1995). Prediction accuracy of the Washington and Illinois risk assessment instruments: An application of receiver operating characteristic curve analysis. *Social Work Research*, *19*(3), 174–183.

Center, C., & Hall, C. (2018). *Making the most of predictive analytics: Responsive and innovative uses in child welfare policy and practice* (tech. rep.). Collaborating at the Intersection of Research and Policy. San Diego, CA & Chicago, IL.

Chaffin, M., & Valle, L. A. (2003). Dynamic prediction characteristics of the child abuse potential inventory. *Child Abuse & Neglect*, *27*(5), 463–481.

Chong, L. E. (1987). Castigo físico e indicadores emocionales en niños: Un estudio a través del dibujo de la figura humana. *Anales de Salud Mental*, *3*, 163–178.

Choo, W. Y., Walsh, K., Chinna, K., & Tey, N. P. (2013). Teacher reporting attitudes scale (TRAS): Confirmatory and exploratory factor analyses with a Malaysian sample. *Journal of Interpersonal Violence*, *28*, 231–253.

Colombo, R. I., & Beigbeder de Agosta, C. (2005). *Abuso y maltrato infantil: Hora de juego diagnóstica*. Ed. Cauquén.

Colombo, R. I., Beigbeder de Agosta, C., & Barilari, Z. (2013). *Indicadores de abuso y maltrato en la prueba gráfica persona bajo la lluvia* (4th ed.). Cauquen Editora.

Coohey, C., Johnson, K., Renner, L. M., & Easton, S. D. (2013). Actuarial risk assessment in child protective services: Construction methodology and performance criteria. *Children and Youth Services Review*, *35*(1), 151–161. http://refhub.elsevier.com/S0145-2134(17)30352-6/sbref0110

Corman, L. (1967). *El test del dibujo de la familia*. Kapelusz.

Dankert, E. W., & Johnson, K. (2014). *Risk assessment validation: A prospective study* (tech. rep.). NCCD Children's Research Center. http://www.nccdglobal.org/sites/default/files/publication_pdf/risk-assessment-validation.pdf

Dinehart, L., & Kenny, M. (2015). Knowledge of child abuse and reporting practices among early care and education providers. *Journal of Research in Childhood Education*, *29*, 429–443.

Elkisch, P. (1966). Expresión artística libre. In *Técnicas proyectivas para niños*. Paidós.

Engstrom, D. F., Ho, D. E., Sharkey, C. M., & Cuéllar, M.-F. (2020). Government by algorithm: Artificial intelligence in federal administrative agencies. (Public Law Research Paper No. 20-54). https://ssrn.com/abstract=3551505

Fayez, M., Takash, H., & Al-Zboon, E. (2014). Combating violence against children: Jordanian preservice early childhood teachers' perceptions towards child abuse and neglect. *Early Child Development and Care*, *184*, 1485–1498.

Feng, J., Huang, T., & Wang, C. (2010). Kindergarten teachers' experience with reporting child abuse in Taiwan. *Child Abuse & Neglect*, *34*(2), 124–128. https://doi.org/https://doi.org/10.1016/j.chiabu.2009.05.007

Fernandes, D., & colleagues. (2024). Forensic interview techniques in child sexual abuse cases: A scoping review [PMC summary available]. *Trauma, Violence, & Abuse*. https://pmc.ncbi.nlm.nih.gov/articles/PMC10913353/

Flaherty, C. W. (2001). *An artificial neural network model for the prediction of child physical abuse recurrences* [Doctoral dissertation, The University of Tennessee]. https://refhub.elsevier.com/S0145-2134(17)30352-6/sbref0205

Flahive, M.-h. W., Chuang, Y.-C., & Li, C.-M. (2015). The multimedia piers-harris children's self-concept scale 2: Its psychometric properties, equivalence with the paper-and-pencil version, and respondent preferences. *Plos one*, *10*(8), e0135386.

Freud, S. (1920). *Más allá del principio de placer* (L. Ballesteros, Trans.). Ed. Biblioteca Nueva.

Garb, H. N., Wood, J. M., & Nezworski, M. T. (2000). Projective techniques and the detection of child sexual abuse. *Child maltreatment*, *5*(2), 161–168.

Gilbert, R., Widom, C. S., Browne, K., Fergusson, D., Webb, E., & Janson, S. (2009). Burden and consequences of child maltreatment in high-income countries. *The Lancet*, *373*(9657), 68–81.

Gobierno de la Ciudad Autónoma de Buenos Aires. (2021). *Maltrato infanto juvenil: Guía de prevención* (tech. rep.). Ministerio de Educación. Ciudad Autónoma de Buenos Aires, Argentina.

Gobierno de la Provincia de Buenos Aires. (2019). *Protocolo provincial de prevención, detección y abordaje del abuso sexual en la infancia* (tech. rep.). Ministerio de Desarrollo Social. Buenos Aires, Argentina. http://www.mds.gba.gov.ar/sai/sai.pdf

Gobierno de la Provincia de Buenos Aires. (2022). *Guía práctica de abordaje integral ante situaciones de violencia sexual contra niñeces y adolescencias* (tech. rep.). Ministerio de Salud. Buenos Aires, Argentina. https://www.ms.gba.gov.ar/sitios/saludmental/files/2022/12/guia_ninez_merged.pdf

Gobierno de la Provincia de La Pampa. (2017). *Protocolo para la intervención en situaciones de abuso sexual contra NNyA* (tech. rep.). Consejo Provincial de la Niñez y Adolescencia. La Pampa, Argentina. https://mds.lapampa.gob.ar/images/Archivos/antePie/Protocolo_ASI_La_Pampa-2017.pdf

Gómez, C. (2015). *Indicadores proyectivos de abuso y maltrato en pruebas psicológicas*. Editorial Morata.

Gómez, C. (2018). Indicadores proyectivos de maltrato infantil en el CAT: Una revisión. *Revista de Psicología Clínica Infantil*, *26*(2), 78–92.

Haj-Yahia, M. M., & Attar-Schwartz, S. (2008). Attitudes of Palestinian pre-school teachers' from Israel towards reporting of suspected cases of child abuse and neglect. *Child and Family Social Work*, *13*, 378–390.

Hammer, E. F. (1958). *The clinical application of projective drawings*. Charles C Thomas.

Hernández, A. (2019). *El test desiderativo en la detección de trauma y maltrato: Perspectivas clínicas*. Editorial Siglo XXI.

Hughes, K., Bellis, M. A., Hardcastle, K. A., Sethi, D., Butchart, A., Mikton, C., Jones, L., & Dunne, M. P. (2017). The effect of multiple adverse childhood experiences on health: A systematic review and meta-analysis. *The Lancet public health*, *2*(8), e356–e366.

ICBF. (2022, June). *Modelos de probabilidad de vulneración para la prevención de violencia contra niñas, niños y adolescentes en colombia* [Instituto Colombiano de Bienestar Familiar. Edición institucional.].

Jaroenkajornkij, N., Lev-Wiesel, R., & Binson, B. (2022). Use of self-figure drawing as an assessment tool for child abuse: Differentiating between sexual, physical, and emotional abuse. *Children*, *9*(6), 868.

Jernbro, C., Otterman, G., Lucas, S., Tindberg, Y., & Janson, S. (2017). Disclosure of child physical abuse and perceived adult support among Swedish adolescents. *Child Abuse Review*, *26*(6), 451–464.

Johnson, W. L. (2011). The validity and utility of the California family risk assessment under practice conditions in the field: A prospective study. *Child Abuse & Neglect*, *35*(1), 18–28.

Johnson, W. L., Clancy, T., & Bastian, P. (2015). Child abuse/neglect risk assessment under field practice conditions: Tests of external and temporal validity and comparison with heart disease prediction. *Children and Youth Services Review*, *56*, 76–87.

Karadag, S. C., Sonmez, S., & Dereobali, N. (2015). An investigation of preschool teachers' recognition of possible child abuse and neglect in Izmir, Turkey. *Journal of Interpersonal Violence*, *30*, 873–891.

Klika, J., Haboush-Deloye, A., & Linkenbach, J. (2019). Hidden protections: Identifying social norms associated with child abuse, sexual abuse, and neglect. *Child and Adolescent Social Work Journal*, *36*. https://doi.org/10.1007/s10560-018-0595-8

Lavoie, J., Cyr, M., Biron, C., et al. (2021). Meta-analysis of the effects of rapport practices and question type on children's disclosures of sensitive information [Open-access summary available via CALiO]. *Child Abuse & Neglect*. https://files.calio.org/Other/Meta-analysis%20of%20the%20effects%20of%20two%20interviewing%20practices%20on%20children%E2%80%99s%20disclosures%20of%20sensitive%20information-%20Rapport%20practices%20and%20question%20type.pdf

Leeuwestein, H., Kupers, E., Boelhouwer, M., Tondera, P., & van Dijk, M. (2024). A screening instrument for trauma-related behavior among young primary school students: Development and validation of the raptoss. *School Mental Health*, *16*, 530–549. https://doi.org/10.1007/s12310-024-09656-8

León Vásquez, R. L., & Castañeda Chang, A. M. (2012). *Indicadores emocionales del test del dibujo de la figura humana de koppitz en niños maltratados y no maltratados* [Tesis de bachiller; disponible en línea], Universidad Nacional Mayor de San Marcos, Lima, Perú. http://www.udd.cl/wp-content/uploads/2012/05/7162191-Test-Indicadores-Emocionales-Del-Test-Del-Dibujo-de-La-Figura-Humana-de-Koppitz-en-NiNos-dos-y-No-dos.pdf

López, M. C. (2014). *Los juegos en la detección del abuso sexual infantil*. Ed. Maipué.

López, M. C. (2015). *La detección del abuso infantil en el ámbito escolar*. Editorial Cinca.

López, M. C. (2017). *La entrevista psicológica a niñas y niños víctimas de abusos sexuales*. Ed. Maipué.

López, M. C. (2020). *Test desiderativo y maltrato infantil: Indicadores proyectivos y métodos de análisis*. Editorial Cinca.

Martínez, M. (2014). *Maltrato infantil: Psicología y prevención*. Editorial Plenitud.

Martinez Sagasti, I. (2020). Desprotección a la infancia en el ámbito familiar: Protocolos de prevención y actuación en materia de protección infantil desde el ámbito escolar. *Familia. Revista De Ciencia Y Orientación Familiar*, (58), 81–98. https://doi.org/10.36576/summa.131284

McTavish, J. R., Gonzalez, A., Santesso, N., MacGregor, J. C. D., McKee, C., & MacMillan, H. L. (2020). Identifying children exposed to maltreatment: A systematic review update. *BMC Pediatrics*, *20*(113). https://doi.org/10.1186/s12887-020-2015-4

Melamed, D. M., Botting, J., Lofthouse, K., Pass, L., & Meiser-Stedman, R. (2024). The relationship between negative self-concept, trauma, and maltreatment in children and adolescents: A meta-analysis. *Clinical child and family psychology review*, *27*(1), 220–234.

Milner, J. S., Gold, R. G., Ayoub, C., & Jacewitz, M. M. (1984). Predictive validity of the child abuse potential inventory. *Journal of Consulting and Clinical Psychology*, *52*(5), 879.

Morín Díaz, M. A., & Ramírez Figallo, N. C. (2017, June). *Análisis psicométrico de los indicadores de maltrato infantil en el test de la casa-árbol-persona en niños de 6 a 12 años* [Trabajo de investigación de grado (Licenciatura en Psicología)], Universidad Católica Andrés Bello, Escuela de Psicología.

Mullen, P. E., Martin, J. L., Anderson, J. C., Romans, S. E., & Herbison, G. P. (1996). The long-term impact of the physical, emotional, and sexual abuse of children: A community study. *Child Abuse & Neglect*, *20*(1), 7–21.

Müller, M. B., & López, M. C. (2011). *Los dibujos en el abuso sexual infantil*. Ed. Maipue.

Münsterberg Koppitz, E. (2006). *El dibujo de la figura humana en los niños* (12a ed. 6a reimp.). Guadalupe.

Murphy, S., Orkow, B., & Nicola, R. M. (1985). Prenatal prediction of child abuse and neglect: A prospective study. *Child Abuse & Neglect*, *9*(2), 225–235.

National Institute for Health and Care Excellence (NICE). (2017). Child abuse and neglect: Recognition and response (ng76) [Clinical guideline NG76].

Norman, R. E., Byambaa, M., De, R., Butchart, A., Scott, J., & Vos, T. (2012). The long-term health consequences of child physical abuse, emotional abuse, and neglect: A systematic review and meta-analysis. *PLoS Medicine*, *9*(11), e1001349.

Oates, R. K., Forrest, D., & Peacock, A. (1985). Self-esteem of abused children. *Child Abuse & Neglect*, *9*(2), 159–163.

Ondersma, S. J., Chaffin, M. J., Mullins, S. M., & LeBreton, J. M. (2005). A brief form of the child abuse potential inventory: Development and validation. *Journal of Clinical Child and Adolescent Psychology*, *34*(2), 301–311.

Pérez de Albéniz Iturriaga, A., Molina, B. L., & Sufrate, M. T. P. (2011). El papel del maestro y la escuela en la protección infantil: Detección de casos y notificación en los servicios de protección infantil en la rioja. *Contextos Educativos. Revista de Educación*, (14), 85–100.

Piers, E. V., & Harris, D. B. (1964). Age and other correlates of self-concept in children. *Journal of Educational Psychology*, *55*(2), 91–95.

Piers, E. V., & Herzberg, D. S. (2002). *Piers-harris 2: Piers-harris children's self-concept scale*. Western Psychological Services.

Pryce, J., Lee, W., Crowe, E., Park, D., McCarthy, M., & Owens, G. (2019). A case study in public child welfare: County-level practices that address racial disparity in foster care placement. *Journal of public child welfare*, *13*(1), 35–59.

Rodulfo, M. (1992). *El niño del dibujo: Estudio psicoanalítico del grafismo y sus funciones en la construcción temprana del cuerpo*. Ed. Paidós.

Saigh, P. A., Yasik, A. E., Oberfield, R., & Halamandaris, P. V. (2008). The self-concept of traumatized children and adolescents with or without ptsd. *Behaviour Research and Therapy*, *46*(10), 1181–1186.

Sánchez, R. (2012). *El test de apercepción temática (CAT) y su aplicación en el diagnóstico del abuso infantil*. Editorial Paidós.

Sánchez, R. (2020). *El test desiderativo en la evaluación del abuso y maltrato infantil: Indicadores y métodos de interpretación*. Editorial Alianza.

Serrano, S., & Villalba, S. (2018). *Dependencia y abuso: Estrategias de intervención para profesionales*. Editorial Alianza.

Sidders, M. M. (2023). *Estudio exploratorio sobre herramientas para medir riesgo de maltrato y abuso en contra de niños y niñas* [Unpublished manuscript].

Sledjeski, E. M., Dierker, L. C., Brigham, R. B., & Breslin, E. (2008). The use of risk assessment to predict recurrent maltreatment: A classification and regression tree analysis (CART). *Prevention Science*, *9*(1), 28–37. https://doi.org/10.1007/s11121-007-0070-9

Stoltenborgh, M., Bakermans-Kranenburg, M. J., & van Ijzendoorn, M. H. (2013). The neglect of child neglect: A meta-analytic review of the prevalence of neglect. *Social psychiatry and psychiatric epidemiology*, *48*(3), 345–355.

Stoltenborgh, M., van Ijzendoorn, M. H., Euser, E. M., & Bakermans-Kranenburg, M. J. (2011). A global perspective on child sexual abuse: Meta-analysis of prevalence around the world. *Child maltreatment*, *16*(2), 79–101.

Svensson, B., & Janson, S. (2008). Suspected child maltreatment: Preschool staff in a conflict of loyalty. *Early Childhood Education Journal*, *36*, 25–31.

Toros, K., & Tiirik, R. (2016). Preschool teachers' perceptions about and experience with child abuse and neglect. *Early Childhood Education Journal*, *44*, 21–30.

UNICEF. (2022). A statistical profile of violence against children in latin america and the caribbean ["Nearly two in three children aged 1 to 14 in Latin America and the Caribbean experience violent discipline at home"]. https://www.unicef.org/lac/reports/statistical-profile-violence-against-children-latin-america-and-caribbean

Vaithianathan, R., Maloney, T., Putnam-Hornstein, E., & Jiang, N. (2013). Children in the public benefit system at risk of maltreatment: Identification via predictive modeling. *American Journal of Preventive Medicine*, *45*(3), 354–359.

van der Put, C. E., Assink, M., Gubbels, J., & van Solinge, N. F. B. (2017). Predicting child maltreatment: A meta-analysis of the predictive validity of risk assessment instruments. *Child Abuse & Neglect*, *73*, 71–88.

van der Put, C. E., Hermanns, J., & van Rijn-van Gelderen, L. (2016). Detection of unsafety in families with parental and/or child developmental problems at the start of family support. *BMC Psychiatry*, *16*(1), 1–12.

Veltman, M. W., & Browne, K. D. (2002). The assessment of drawings from children who have been maltreated: A systematic review. *Child Abuse Review: Journal of the British Association for the Study and Prevention of Child Abuse and Neglect*, *11*(1), 19–37.

Walsh, K., Zwi, K., Woolfenden, S., Shlonsky, A., El-Murr, A., Gibbs, L., Lonne, B., & Higgins, D. J. (2022). Child protection training for professionals to improve reporting of child abuse and neglect. *Cochrane Database of Systematic Reviews*, (7), CD011775. https://doi.org/10.1002/14651858.CD011775.pub2

WHO. (2024, November). Child Maltreatment [Accessed: 2025-07-02]. https://www.who.int/news-room/fact-sheets/detail/child-maltreatment

Wood, S. M. (1997). Risk predictors for re-abuse or re-neglect in a predominantly hispanic population. *Child Abuse & Neglect*, *21*(4), 379–389.

Zhang, H., Wang, W., Liu, S., Feng, Y., & Wei, Q. (2023). A meta-analytic review of the impact of child maltreatment on self-esteem: 1981 to 2021. *Trauma, Violence, & Abuse*, *24*(5), 3398–3411.

Annex 1. Overview of Risk Assessment Tools for Child Maltreatment

This annex presents a summary of the 18 risk assessment tools reviewed as part of the literature analysis. These tools vary in terms of their target populations, data sources, and intended applications. The majority focus on families already involved with child protection systems or on high-risk groups, with relatively few designed for use in general population settings such as schools. The table below indicates the type of input data required by each tool and the main reference sources consulted in this review.

Tool	Population	Type of input data	Reference
Connecticut Risk Assessment-CART model	Children in the child protection system	Administrative data	Sledjeski et al. (2008)
Connecticut Risk Assessment-regression model	Children in the child protection system	Administrative data	Sledjeski et al. (2008)
Modelos de probabilidad de vulneración para la prevención (Colombia)	At-risk families	Administrative data	ICBF (2022)
Neural network model	General population	Administrative data	Flaherty, C. W. (2001)
Predictive Risk Model (V1)	At-risk families	Administrative data	Vaithianathan et al. (2013)
Severe Harm Predictive Model (New York - USA)	Children in the child protection system	Direct observation / survey	Engstrom et al. (2020)
Sistema de Alerta Niñez (Chile)	At-risk families	Administrative data	Algoritmos Públicos (2021)
California Family Risk Assessment (CFRA)	Children in the child protection system	Direct observation / survey	Johnson (2011), Johnson et al. (2015), Dankert & Johnson (2014), van der Put et al. (2016)
Child Abuse Potential Inventory (CAPI)	Children in the child protection system	Direct observation / survey	Milner et al. (1984), Ayoub & Milner (1985), Chaffin & Valle (2003), Ondersma et al. (2005)
Colorado Family Risk Assessment Abuse scale	Children in the child protection system	Direct observation / survey	Coohey et al. (2013)
Colorado Family Risk Assessment Revised - Abuse scale	Children in the child protection system	Direct observation / survey	Coohey et al. (2013)
Family Stress Checklist	General population	Direct observation / survey	Murphy et al. (1985)
Maternal History Interview (MHI)	At-risk families	Direct observation / survey	Altemeier et al. (1984), Brayden et al. (1993)
NCCD Risk Assessment Tools	General population	Administrative data	Wood (1997)
Y-ACNAT-NO	At-risk families	Mixed data	Assink et al. (2015)
CANTS 17B	Children in the child protection system	Mixed data	Camasso & Jagannathan (1995)
Ontario Risk Assessment Tool	Children in the child protection system	Mixed data	Barber et al. (2008)
Washington Risk Assessment Matrix (WRAM)	Children in the child protection system	Mixed data	Camasso & Jagannathan (1995), Baird & Wagner (2000)

Annex 2. Graphic Indicators Used in the Analysis of Projective Drawings

This annex presents the set of graphic indicators used in the analysis of children's projective drawings within the study. These indicators were applied to evaluate the Kinetic Family Drawing and Person Under the Rain tasks, and were used to identify patterns that may suggest experiences of abuse or neglect.

Kinetic Family

- **Poor overall impression of the drawing:** the drawing is not visually pleasing, it looks "weird" or ominous.
- **Poor integration:** either in the overall drawing or in the body, elements appear "scattered", not generating cohesion in the graphic production. This indicator is valid from 6 years old in girls and 7 years old in boys.
- **Shaded face:** the face of at least one member appears shadowed, or darkened, making it difficult to see their features.
- **Arms glued to the body / legs together:** there is no separation between the external body parts, they appear as if they were "sealed". Valid from 5 years old in girls and from 6 years old in boys.
- **Strange suns:** large suns are observed with bizarre features.
- **Omission of eyes:** at least one human figure is drawn without eyes.
- **Dark colors:** painted with colors like black, brown, red, gray, or dull gold colors.
- **Unpleasant expression of characters:** characters appear sad, with a disgusted or uncomfortable face, or in a strange pose. They can also be observed with teeth, bulging eyes, or frightening facial expressions.
- **Amorphous figure / distortion of body schema:** disintegrated bodies, with a strange, non-human schema.
- **Dismembered figures:** body elements are observed disintegrated, most commonly at the level of arms and legs with the body and/or at the level of the belly or head, showing them separated from the rest of the body.
- **Empty eyes / without pupils:** draws the eye socket without pupils. Valid in boys and girls from 6 years old onwards.
- **Omission of the mouth:** at least one character does not have a mouth.
- **Omission of arms:** at least one character does not have limbs or upper limbs. Valid in girls from 5 years old and in boys from 6 years old.
- **Family characters with marks in the genital area:** marks, details, or scribbles are observed in the genital areas.

- **Crossed-out family characters:** at least one character is crossed out, scribbled over, or painted with a color in a way that details of the figure cannot be observed.
- **Interrupted and angled stroke:** lines are not clean, or with marked angles (excluding Minecraft-style characters, common in today's childhoods).
- **Absence of family:** draws other content in which a family group is not observed (can be a single person as well).
- **Naked people:** at least one individual is not wearing clothes (may appear "white", unpainted figure, empty).

Person under the Rain

- **Amorphous figure:** the schema of the human figure does not match the expected one for its evolutionary development.
- **Interrupted and angled stroke:** lines are not clean, or with marked angles (excluding Minecraft-style characters, common in today's childhoods).
- **Detail in genital area:** marks, elements, or scribbles are observed in the genital areas.
- **Absence of feet:** at least one figure lacks feet (not legs). This indicator is not significant until 7 years old in girls and 9 in boys.
- **Childish or incomplete figure:** drawings appear more childish or clumsy than expected for their age, making legs, arms, fingers, etc., of a character with one-dimensional lines.
- **Divided and double figure:** the human figure appears split, double, or behind it, another figure can be seen that may be poorly erased.
- **Phallic elements:** presents elements, details, or figures similar to phalluses.
- **Body and stroke rigidity:** figures are observed as static, immobilized, without expression of life, like statues. Valid in girls and boys from 6 years old onwards.
- **Absence of a body part:** the human figure is not complete or appears disintegrated due to the lack of one of its parts.
- **Poor overall appearance:** the drawing is not visually pleasing, it looks "weird" or ominous.
- **Special treatment of feet:** presents details, colors, or unusual elements in the feet or footwear (for example, yellow color, boots, etc.).
- **Naked figure:** at least one individual is not wearing clothes (may appear "white," unpainted figure, empty).

For the general evaluation of the drawing and the identification of "high" or "low" in the presence of indicators, it is recommended to be guided primarily by the overall impression and the presence or absence of serious indicators such as naked figures or phallic elements in the drawings.

Annex 3. Scoring and Interpretation of the Piers-Harris Self-Concept Scale

To evaluate self-concept, the standard correction template of the Piers-Harris Self-Concept Scale was applied, assigning one point for each response that corresponds to the marked symbol. Each symbol on the template is associated with one of the six dimensions measured by the scale, with the exception of items 7 and 12, which contribute to two dimensions simultaneously. Based on the responses, raw scores are obtained for each dimension, which are then converted to percentile scores using reference benchmarks specific to school level and country or cultural context. These percentiles are reflected in the individual results sheet. The sum of the scores across all dimensions produces the Global Self-Concept score. According to the benchmarks applied, scores above the 65th percentile are classified as indicative of good self-concept, while scores below the 35th percentile indicate poor or diminished self-concept. The results sheet also includes a section where the tutor or teacher can indicate whether the results are consistent with their own observations of the child.

#	STATEMENT			#	STATEMENT		
1	My classmates make fun of me	Yes	No	41	I have beautiful hair	Yes	No
2	I am a happy person	Yes	No	42	I often volunteer in class	Yes	No
3	I find it difficult to make friends	Yes	No	43	I would like to be different from how I am	Yes	No
4	I am sad often	Yes	No	44	I sleep well at night	Yes	No
5	I am smart	Yes	No	45	I hate school	Yes	No
6	I am shy	Yes	No	46	I am usually chosen last to play	Yes	No
7	I get nervous when the teacher asks	Yes	No	47	I am often sick	Yes	No
8	I dislike my face	Yes	No	48	I am often unfriendly with others	Yes	No
9	When I grow up, I will be an important person	Yes	No	49	My peers think I have good ideas	Yes	No
10	I worry a lot when we have a test	Yes	No	50	I am unhappy	Yes	No
11	I am unpopular in class	Yes	No	51	I have many friends	Yes	No
12	I misbehave in class	Yes	No	52	I am cheerful	Yes	No
13	When something goes wrong, it's usually my fault	Yes	No	53	I am clumsy in most things	Yes	No
14	I cause problems for my family	Yes	No	54	I am handsome/beautiful	Yes	No
15	I am strong	Yes	No	55	When I have to do something, I do it enthusiastically	Yes	No
16	I have good ideas	Yes	No	56	I fight a lot	Yes	No
17	I am an important member of my family	Yes	No	57	Girls like me	Yes	No
18	I usually want my own way	Yes	No	58	People take advantage of me	Yes	No
19	I am good with my hands	Yes	No	59	My family is disappointed with me	Yes	No
20	When things are difficult, I leave them undone	Yes	No	60	I have a pleasant face	Yes	No
21	I do my schoolwork well	Yes	No	61	When I try to do something, it seems to go wrong	Yes	No
22	I do many bad things	Yes	No	62	My family takes advantage of me	Yes	No
23	I draw well	Yes	No	63	I am one of the best in games and sports	Yes	No
24	I am good at music	Yes	No	64	I am clumsy	Yes	No
25	I misbehave at home	Yes	No	65	In games and sports, I watch but do not participate	Yes	No
26	I am slow at doing my schoolwork	Yes	No	66	I forget what I learn	Yes	No
27	I am an important member of my class	Yes	No	67	I get along with people	Yes	No
28	I am nervous	Yes	No	68	I get angry easily	Yes	No
29	I have pretty eyes	Yes	No	69	Boys like me	Yes	No
30	I can make a good impression in class	Yes	No	70	I read well	Yes	No
31	I daydream in class	Yes	No	71	I prefer to work alone rather than in a group	Yes	No
32	I annoy my siblings	Yes	No	72	I get along with my siblings	Yes	No
33	My friends like my ideas	Yes	No	73	I have a good body	Yes	No
34	I often get into trouble	Yes	No	74	I often feel afraid	Yes	No
35	I am obedient at home	Yes	No	75	I'm always breaking things	Yes	No
36	I am lucky	Yes	No	76	I can be trusted	Yes	No
37	I worry a lot about things	Yes	No	77	I'm a weird person	Yes	No
38	My parents demand too much of me	Yes	No	78	I think about doing bad things	Yes	No
39	I like being who I am	Yes	No	79	I cry easily	Yes	No
40	I feel a little rejected	Yes	No	80	I am a good person	Yes	No

Annex 4. Teacher's Checklist for Indicators of Child Maltreatment and Neglect

This annex presents the complete checklist used by teachers in the study to identify potential indicators of child maltreatment and neglect. The checklist consists of 20 indicators, selected and prioritized based on national and provincial guidelines, with an emphasis on relevance and ease of observation in the school context. The tool was designed to support teachers in systematically recording physical, behavioral, psychological, and academic signs that may warrant further assessment.

Indicators and signs of child abuse and/or neglect		
Physical indicators		
"Have you frequently observed bruises, burns, and/or fractures in the child?"	Yes	No
"Have you observed if the child experiences recurrent psychosomatic disorders? (e.g., abdominal pain, headaches, alopecia, asthma attacks, etc.)"	Yes	No
"Have you observed if the child has eating disorders? (obesity, bulimia, anorexia nervosa, eating anxiety)"	Yes	No
"Have you noticed if the child lacks control of their sphincters?"	Yes	No
"Have you noticed if the child experiences recurrent urinary tract infections?"	Yes	No
Behavioural indicators		
"Have you noticed if the child exhibits significant changes in behavior to opposite extremes? (from calm to restless, or vice versa)"	Yes	No
"Have you noticed if the child exhibits any explicit sexual behavior? (inappropriate language, sexual knowledge and curiosity beyond their age, masturbation, sexual play, sharing pornography)"	Yes	No
"Have you observed risk behaviors in the child? (self-destructive)"	Yes	No
"Have you noticed if the child experiences a decline in self-esteem, loss of interest in their surroundings, and/or in the future?"	Yes	No
"Have you noticed if the child has suicidal thoughts?"	Yes	No
"Have you noticed if the child exhibits intense phobias and/or fears?"	Yes	No
"Does the child often experience episodes of excessive crying?"	Yes	No
Indicators in the school life		
"Does the child have repeated absences without reason or with excuses?"	Yes	No
"Have you noticed if the child resists changing clothes for various activities? Do they wear long sleeves/loose clothing or a hood on hot days?"	Yes	No
"Have you noticed neglect of the child's health by the responsible adults? (for example, not taking them to the doctor for a prolonged illness)"	Yes	No
"Have you noticed changes in the child's academic performance?"	Yes	No
"Have you noticed if the child tends to stay at school beyond the usual hours?"	Yes	No
"Does the child engage in running away, vandalism, and/or petty theft?"	Yes	No
Psychological indicators		
"Did the child report that they suffered physical or sexual abuse? Or did any guardian of that child do so?"	Yes	No
Additional comments		
"Additional comments about the child"	Yes	No

In the presence of a physical indicator (such as frequent injuries, burns, and fractures) or more than one indicator of change in behavior or school life, the case is classified as high risk. Likewise, if the Teacher's Checklist includes the disclosure of a child having suffered physical or sexual abuse, these cases must be reported immediately as they are considered high risk.

Annex 5: Sample Size Calculations

We estimate the sample size needed to accurately calculate the precision metric, this is, the proportion of children who, after being flagged by our tools, are confirmed to suffer maltreatment or neglect by the comprehensive psychological assessment (our reference test).

To achieve this, we applied the formula that allows us to estimate the necessary sample size for calculating a population proportion p with a confidence interval of $(1-\alpha)100\%$ and an error no greater than $\varepsilon : N = \frac{z_{\frac{\alpha}{2}}^2 p(1-p)}{\varepsilon^2}$. We determined that to identify any population proportion with a margin of error of less than 6 percentage points, with a 95% confidence interval we would need at least 267 observations.

This implies that by conducting psychological assessments on approximately 270 children, we will achieve sufficient precision to estimate the population proportion of those who, having been identified by the diagnostic tools as high risk, are indeed true positives.

Given the estimated precision of our tools, the need arises to compare how many positive cases are identified among those children who were not identified by the intervention. Essentially, our goal is to test the hypothesis that our predictive tools can accurately differentiate between children at risk of abuse or mistreatment and those who are not. In essence, we aim to validate the following hypothesis:

$$H_0 : p_0 = p_1 \quad \text{vs} \quad H_1 : p_0 \neq p_1$$

In this context, p_1 signifies the proportion of positive cases, as identified through psychological analysis, within the subgroup that triggered a red flag via the diagnostic tools. Conversely, p_0 denotes the proportion of true positives, also ascertained through psychological analysis, among the individuals who did not trigger a red flag according to the screening tools.

To calculate the necessary sample size for the subgroup that did not trigger a red flag, enabling us to confidently reject the null hypothesis, we employ the equation:

$$n_1 = \frac{\left(z_{\frac{\alpha}{2}}\sqrt{(k+1)\underline{p}(1-\underline{p})} + z_\beta\sqrt{p_0(1-p_0)+kp_1(1-p_1)}\right)^2}{k(p_1-p_0)^2}, \quad n_0 = k \cdot n_1$$

In this formulation, n_1 denotes the sample size needed for the subgroup that triggered a red flag, while n_0 specifies the sample size required for the subgroup without a red flag. The variable k represents the ratio of the size of the subgroup without a red flag to the size of the subgroup with a red flag (with a value of 1 indicating equal sizes for both groups). The term $z_{\frac{\alpha}{2}}$ corresponds to the z-score associated with achieving a confidence level of $(1-\alpha)100\%$, such as 1.96 for 95% confidence intervals, indicating the standard deviation from the mean. Meanwhile, z_β refers to the z-score necessary to attain a specified statistical power, with 0.84 corresponding to 80% power, reflecting the standard normal deviation. Finally, p is defined as:

$$\underline{p} = \frac{(kp_0 + p_1)}{k+1}$$

Leveraging an informed estimate that no more than 30% of positive cases would be detected among children not triggering a red flag, we adopted $p_0 = 0.3$ as a plausible value. To err on the side of caution,

we set $p_1 = 0.5$, with a confidence level $\alpha = 0.95$ and a power $\beta = 0.8$, alongside a proposed sample size for the red flagged subgroup $n_1 = 250$. Based on these parameters, our calculations indicate that we require a sample of 58 children without a red flag to undergo psychological evaluation.

After diagnosing this sample of approximately 60 children, we can obtain an estimate of p_0 (the true rate of positive cases without a red flag).

Annex 6. Methodology for Estimating Key Metrics

This annex details the estimation strategy used to recover *prevalence*, *sensitivity*, *specificity*, and *precision* (positive predictive value, PPV) of the early-detection tools when only a subset of children receive the comprehensive psychological assessment (CPA). The algebra extends the outline in Section 5 by deriving large–sample standard errors with the delta method.

A. Notation

- $X_i = 1$ if the screening tool raises a red flag for child i; 0 otherwise.
- $Y_i = 1$ if the CPA confirms maltreatment; 0 otherwise.
- $R_i = 1$ if child i is selected for CPA; 0 otherwise.

Sampling design

$$\pi_i = \Pr(R_i = 1) = \begin{cases} 1 & \text{if } X_i = 1, \\ \tilde{\pi} & \text{if } X_i = 0, \end{cases} \qquad \tilde{\pi} = \frac{\#\text{CPAs among } X = 0}{\#\text{children with } X = 0}.$$

Each observation is weighted by R_i/π_i (inverse probability weighting).

B. Target Parameters

$$\begin{aligned} \lambda &= P(Y = 1) && \text{Prevalence,} \\ \theta &= P(X = 1 \mid Y = 1) && \text{Sensitivity,} \\ \beta &= P(X = 0 \mid Y = 0) && \text{Specificity,} \\ p_1 &= P(Y = 1 \mid X = 1) && \text{Precision (PPV).} \end{aligned}$$

C. Estimating Prevalence

$$\hat{\lambda} = \frac{1}{m} \sum_{i=1}^{m} \frac{R_i}{\pi_i} Y_i.$$

$\hat{\lambda}$ solves $\sum_i \frac{R_i}{\pi_i}(Y_i - \hat{\lambda}) = 0$. Linearising,

$$\sqrt{m}\,(\hat{\lambda} - \lambda) \xrightarrow{d} N\Big(0, \sigma_\lambda^2\Big), \quad \sigma_\lambda^2 = Var\Big(\tfrac{R}{\pi}(Y - \lambda)\Big).$$

$$\hat{\sigma}^2_\lambda = \frac{1}{m}\sum_{i=1}^{m} \frac{R_i}{\pi_i^2}\big(Y_i - \hat{\lambda}\big)^2, \qquad \text{CI}_{95\%}:\ \hat{\lambda} \pm 1.96\,\frac{\sqrt{\hat{\sigma}^2_\lambda}}{\sqrt{m}}.$$

D. Estimating Sensitivity

$$\hat{\theta} = \frac{\sum_i \frac{R_i}{\pi_i} X_i Y_i}{\sum_i \frac{R_i}{\pi_i} Y_i}.$$

Estimating-equation: $\sum_i \frac{R_i}{\pi_i} Y_i\big(X_i - \hat{\theta}\big) = 0$. Hence

$$\sqrt{n}\,(\hat{\theta} - \theta) \xrightarrow{d} N\Big(0, \sigma^2_\theta\Big), \quad \sigma^2_\theta = \frac{\text{Var}\big(\frac{R}{\pi} Y (X - \theta)\big)}{\big[E\big(\frac{R}{\pi} Y\big)\big]^2}.$$

$$\hat{\sigma}^2_\theta = \frac{\frac{1}{n}\sum_i \frac{R_i}{\pi_i^2} Y_i \big(X_i - \hat{\theta}\big)^2}{\Big(\frac{1}{n}\sum_i \frac{R_i}{\pi_i} Y_i\Big)^2}, \qquad \text{CI}_{95\%}:\ \hat{\theta} \pm 1.96\,\frac{\sqrt{\hat{\sigma}^2_\theta}}{\sqrt{n}}.$$

E. Estimating Specificity

$$\hat{\beta} = \frac{\sum_i \frac{R_i}{\pi_i}(1 - X_i)(1 - Y_i)}{\sum_i \frac{R_i}{\pi_i}(1 - Y_i)}.$$

Estimating-equation: $\sum_i \frac{R_i}{\pi_i}(1 - Y_i)\big((1 - X_i) - \hat{\beta}\big) = 0$. Thus

$$\sqrt{n}\,(\hat{\beta} - \beta) \xrightarrow{d} N\Big(0, \sigma^2_\beta\Big), \quad \sigma^2_\beta = \frac{\text{Var}\big(\frac{R}{\pi}(1 - Y)\big((1 - X) - \beta\big)\big)}{\big[E\big(\frac{R}{\pi}(1 - Y)\big)\big]^2}.$$

$$\hat{\sigma}^2_\beta = \frac{\frac{1}{n}\sum_i \frac{R_i}{\pi_i^2}(1 - Y_i)\big((1 - X_i) - \hat{\beta}\big)^2}{\Big(\frac{1}{n}\sum_i \frac{R_i}{\pi_i}(1 - Y_i)\Big)^2}, \qquad \text{CI}_{95\%}:\ \hat{\beta} \pm 1.96\,\frac{\sqrt{\hat{\sigma}^2_\beta}}{\sqrt{n}}.$$

F. Estimating Precision (PPV)

All red-flagged children ($X = 1$) were verified, so weighting is unnecessary:

$$\hat{p}_1 = \frac{\sum_i R_i X_i Y_i}{\sum_i R_i X_i}, \qquad k = \sum_i R_i X_i.$$

The usual binomial variance gives

$$\mathrm{CI}_{95\%}: \ \hat{p}_1 \pm 1.96 \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{k}}.$$

G. Summary of Estimators

Table A.1: *Estimators and large-sample confidence intervals*

Metric	Point estimator	95% CI (delta method)
Prevalence λ	$\hat{\lambda} = \frac{1}{m}\sum \frac{R}{\pi} Y$	$\hat{\lambda} \pm 1.96 \frac{\sqrt{\hat{\sigma}^2_{\hat{\lambda}}}}{\sqrt{m}}$
Sensitivity θ	$\hat{\theta} = \frac{\sum \frac{R}{\pi} XY}{\sum \frac{R}{\pi} Y}$	$\hat{\theta} \pm 1.96 \frac{\sqrt{\hat{\sigma}^2_{\hat{\theta}}}}{\sqrt{n}}$
Specificity β	$\hat{\beta} = \frac{\sum \frac{R}{\pi}(1-X)(1-Y)}{\sum \frac{R}{\pi}(1-Y)}$	$\hat{\beta} \pm 1.96 \frac{\sqrt{\hat{\sigma}^2_{\hat{\beta}}}}{\sqrt{n}}$
Precision p_1	$\hat{p}_1 = \frac{\sum RXY}{\sum RX}$	$\hat{p}_1 \pm 1.96 \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{k}}$

These expressions deliver unbiased point estimates and delta-method standard errors under the study's two-stage verification design.

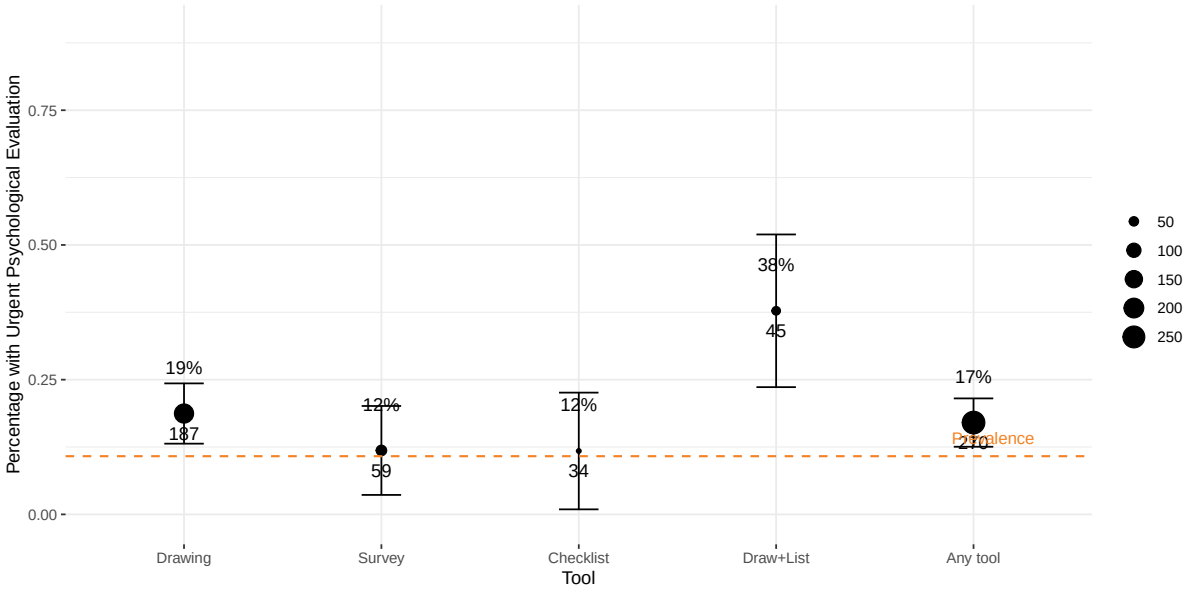
Annex 7. Additional Tables and Figures

Table A.2: *Sample Characteristics*

Characteristic	N (%)
Total sample	1,087
Gender	
Female	539 (49.6)
Male	548 (50.4)
Age	
Mean	8.6
Range	5–14
Age groups	
5–7 years	362 (33.3)
8–9 years	351 (32.3)
10–14 years	374 (34.4)

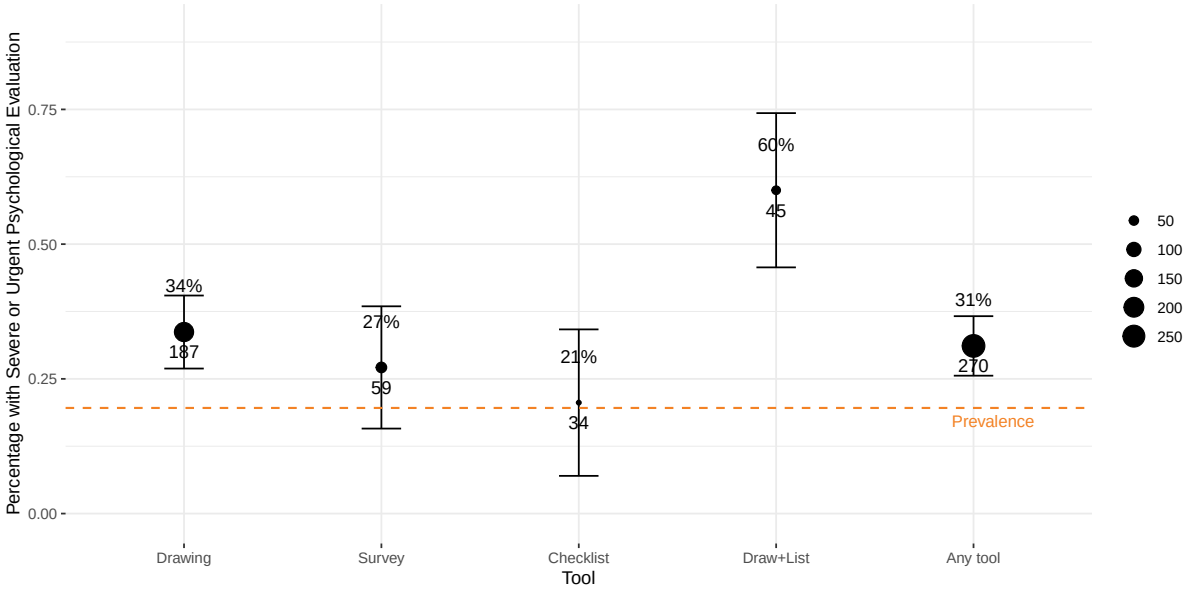
Source: Authors' calculations. **Note:** Participants included children from five primary schools in Buenos Aires City and Córdoba, Argentina.

Figure A.1: *Precision of the Diagnostic Tools for Urgent Maltreatment Cases*



Source: Authors' calculations. **Note:** The dot position shows the percentage of cases confirmed as urgent maltreatment through comprehensive psychological assessments. Dot size represents the number of cases assessed with a red flag for each tool or combination of tools. Whiskers indicate 95% confidence intervals.

Figure A.2: *Precision of the Diagnostic Tools for Severe and Urgent Maltreatment Cases*



Source: Authors' calculations. **Note:** The dot position shows the percentage of cases confirmed as severe or urgent maltreatment through comprehensive psychological assessments. Dot size represents the number of cases assessed with a red flag for each tool or combination of tools. Whiskers indicate 95% confidence intervals.

Table A.3: *Precision and Odds Ratios of Screening-Tool Combinations, by Severity of Maltreatment*

ODDS RATIO			ODDS RATIO			ODDS RATIO		
Positives cases			Severe or Urgent cases			Urgent cases		
TOOL	PRECISION	ODDS	TOOL	PRECISION	ODDS	TOOL	PRECISION	ODDS
Drawings	0.52	1.74	Drawings	0.35	2.23	Drawings	0.20	2.23
Survey	0.54	1.82	Survey	0.28	1.75	Survey	0.12	1.38
Checklist	0.50	1.68	Checklist	0.23	1.48	Checklist	0.13	1.52
Draw + Surv	0.43	1.44	Draw + Surv	0.21	1.36	Draw + Surv	0.14	1.63
Draw + Checklist	0.76	2.53	Draw + Checklist	0.63	3.98	Draw + Checklist	0.40	4.51
Surv + Checklist	0.53	1.79	Surv + Checklist	0.27	1.69	Surv + Checklist	0.13	1.52
Draw + Surv + Checklist	0.50	1.68	Draw + Surv + Checklist	0.17	1.06	Draw + Surv + Checklist	0.17	1.90
Draw—Surv	0.53	1.78	Draw—Surv	0.34	2.16	Draw—Surv	0.18	2.04
Draw—Checklist	0.50	1.68	Draw—Checklist	0.33	2.08	Draw—Checklist	0.19	2.20
Surv—Checklist	0.52	1.75	Surv—Checklist	0.26	1.67	Surv—Checklist	0.13	1.44
Any tool	0.52	1.74	Any tool	0.33	2.07	Any tool	0.18	2.04
Control synth	0.35	1.19	Control synth	0.20	1.24	Control synth	0.11	1.24
No red flag	0.30		No red flag	0.16	1.00	No red flag	0.09	

Source: Authors' calculations. **Note:** Precision represents the percentage of red-flagged cases confirmed as maltreatment through comprehensive psychological assessments or prior school identification. Odds ratios compare the likelihood of confirmed maltreatment among children flagged by the tool (or tool combination) to those without a red flag.

Table A.4: *Results for Any Type of Maltreatment: Prevalence, Confusion Matrix Metrics, and Sample Size by Tool*

	Prevalence	Confusion Matrix				Observations	
Tool		Sensitivity	Specificity	PPV	NPV	Total	With red flag
Any tool	35.3 (26.2-44.4)	36.8 (25.1-48.5)	81.3 (14.3-100.0)	51.9 (45.9-57.8)	70.2 (67.0-73.3)	1077	270
Drawing	35.3 (26.2-44.4)	25.5 (16.9-34.1)	87.1 (29.7-100.0)	51.9 (44.7-59.0)	68.1 (65.1-71.2)	1077	187
Checklist	39.6 (20.3-58.8)	13.1 (3.4-22.8)	91.5 (30.4-100.0)	50.0 (33.2-66.8)	61.6 (56.1-67.2)	329	34
Survey	28.9 (14.3-43.6)	30.1 (11.5-48.8)	89.6 (5.1-100.0)	54.2 (41.5-66.9)	75.9 (71.1-80.7)	367	59
Draw+Surv	28.9 (14.3-43.6)	5.7 (0.3-11.0)	96.9 (51.5-100.0)	42.9 (16.9-68.8)	71.6 (66.9-76.3)	367	14
Draw+List	35.7 (25.6-45.8)	15.3 (8.8-21.9)	97.2 (47.1-100.0)	75.6 (63.0-88.1)	67.4 (63.6-71.2)	620	45
Draw+Surv+List	26.6 (12.6-40.5)	5.1 (0.0-11.3)	98.1 (49.6-100.0)	50.0 (10.0-90.0)	74.1 (68.2-79.9)	220	6
Surv+List	26.6 (12.6-40.5)	13.7 (2.8-24.6)	95.7 (31.0-100.0)	53.3 (28.1-78.6)	75.4 (69.5-81.3)	220	15
Draw—Surv	35.3 (26.2-44.4)	32.3 (21.8-42.8)	84.3 (20.9-100.0)	53.0 (46.6-59.4)	69.5 (66.4-72.6)	1077	232
Draw—List	35.3 (26.2-44.4)	31.3 (21.1-41.4)	83.5 (20.9-100.0)	50.9 (44.4-57.3)	69.0 (65.8-72.1)	1077	234
Surv—List	35.5 (21.9-49.2)	23.3 (11.3-35.2)	88.3 (20.4-100.0)	52.2 (42.0-62.4)	67.6 (63.5-71.8)	581	92

Source: Authors' calculations. **Note:** Prevalence, sensitivity, and specificity include 95% confidence intervals derived as shown in Annex 6. Precision (PPV) represents the percentage of red-flagged cases confirmed as maltreatment through comprehensive psychological assessments or prior school identification.

Table A.5: *Results for Severe and Urgent Maltreatment: Prevalence, Confusion Matrix Metrics, and Sample Size by Tool*

	Prevalence	Confusion Matrix				Observations	
Tool		Sensitivity	Specificity	PPV	NPV	Total	With red flag
Any tool	19.6 (12.4-26.9)	39.7 (23.3-56.2)	78.5 (0.0-100.0)	31.1 (25.6-36.6)	84.2 (81.7-86.7)	1077	270
Drawing	19.6 (12.4-26.9)	29.8 (16.9-42.7)	85.7 (15.3-100.0)	33.7 (26.9-40.5)	83.3 (80.9-85.8)	1077	187
Checklist	15.9 (1.6-30.2)	13.3 (0.0-28.9)	90.2 (13.0-100.0)	20.6 (7.0-34.2)	84.6 (80.5-88.7)	329	34
Survey	14.3 (3.2-25.4)	30.4 (3.7-57.2)	86.3 (0.0-100.0)	27.1 (15.8-38.5)	88.1 (84.5-91.7)	367	59
Draw+Surv	14.3 (3.2-25.4)	5.7 (0.0-13.4)	96.5 (42.2-100.0)	21.4 (0.0-42.9)	86.0 (82.3-89.6)	367	14
Draw+List	18.7 (11.4-26.1)	23.2 (11.8-34.7)	96.4 (30.0-100.0)	60.0 (45.7-74.3)	84.5 (81.5-87.4)	620	45
Draw+Surv+List	9.5 (4.4-14.7)	4.8 (0.0-13.9)	97.5 (38.3-100.0)	16.7 (0.0-46.5)	90.7 (86.8-94.6)	220	6
Surv+List	9.5 (4.4-14.7)	19.0 (2.3-35.8)	94.5 (14.1-100.0)	26.7 (4.3-49.0)	91.7 (87.9-95.5)	220	15
Draw—Surv	19.6 (12.4-26.9)	35.9 (20.8-51.0)	82.0 (5.8-100.0)	32.8 (26.7-38.8)	84.0 (81.5-86.4)	1077	232
Draw—List	19.6 (12.4-26.9)	35.0 (20.2-49.8)	81.5 (6.2-100.0)	31.6 (25.7-37.6)	83.7 (81.2-86.2)	1077	234
Surv—List	17.4 (6.6-28.2)	22.7 (5.8-39.6)	85.6 (5.1-100.0)	25.0 (16.2-33.8)	84.0 (80.8-87.3)	581	92

Source: Authors' calculations. **Note:** Prevalence, sensitivity, and specificity include 95% confidence intervals derived as shown in Annex 6. Precision (PPV) represents the percentage of red-flagged cases confirmed as severe and urgent maltreatment through comprehensive psychological assessments.

Table A.6: *Results for Urgent Maltreatment: Prevalence, Confusion Matrix Metrics, and Sample Size by Tool*

	Prevalence	Confusion Matrix				Observations	
Tool		Sensitivity	Specificity	PPV	NPV	Total	With red flag
Any tool	10.8 (5.2-16.5)	39.4 (17.4-61.4)	76.7 (0.0-100.0)	17.0 (12.6-21.5)	91.2 (89.3-93.2)	1077	270
Drawing	10.8 (5.2-16.5)	30.0 (12.5-47.4)	84.2 (2.3-100.0)	18.7 (13.1-24.3)	90.8 (88.9-92.7)	1077	187
Checklist	13.5 (0.0-27.8)	9.0 (0.0-22.0)	89.5 (18.7-100.0)	11.8 (0.9-22.6)	86.3 (82.4-90.2)	329	34
Survey	2.7 (0.8-4.6)	70.0 (41.6-98.4)	85.4 (0.0-100.0)	11.9 (3.6-20.1)	99.0 (97.9-100.0)	367	59
Draw+Surv	2.7 (0.8-4.6)	20.0 (0.0-44.8)	96.6 (0.0-100.0)	14.3 (0.0-32.6)	97.7 (96.2-99.3)	367	14
Draw+List	13.4 (6.4-20.5)	20.4 (6.8-34.0)	94.8 (22.5-100.0)	37.8 (23.6-51.9)	88.5 (85.9-91.1)	620	45
Draw+Surv+List	4.5 (1.4-7.7)	10.0 (0.0-28.6)	97.6 (13.1-100.0)	16.7 (0.0-46.5)	95.8 (93.1-98.5)	220	6
Surv+List	4.5 (1.4-7.7)	20.0 (0.0-44.8)	93.8 (0.0-100.0)	13.3 (0.0-30.5)	96.1 (93.4-98.7)	220	15
Draw—Surv	10.8 (5.2-16.5)	34.2 (14.7-53.8)	80.0 (0.0-100.0)	17.2 (12.4-22.1)	90.9 (89.0-92.9)	1077	232
Draw—List	10.8 (5.2-16.5)	36.0 (15.6-56.3)	80.0 (0.0-100.0)	17.9 (13.0-22.9)	91.1 (89.2-93.0)	1077	234
Surv—List	8.7 (0.7-16.7)	21.7 (0.0-44.8)	84.7 (0.0-100.0)	12.0 (5.3-18.6)	91.9 (89.5-94.3)	581	92

Source: Authors' calculations. **Note:** Prevalence, sensitivity, and specificity include 95% confidence intervals derived as shown in Annex 6. Precision (PPV) represents the percentage of red-flagged cases confirmed as urgent maltreatment through comprehensive psychological assessments.

Table A.7: *Results for Negligence: Prevalence, Confusion Matrix Metrics, and Sample Size by Tool*

	Prevalence	Confusion Matrix				Observations	
Tool		Sensitivity	Specificity	PPV	NPV	Total	With red flag
Any tool	20.8 (12.5-29.2)	38.2 (20.9-55.5)	81.3 (0.0-100.0)	35.0 (28.4-41.6)	83.3 (80.5-86.1)	880	200
Drawing	20.8 (12.5-29.2)	25.1 (13.0-37.2)	87.1 (18.8-100.0)	33.8 (25.9-41.8)	81.5 (78.8-84.3)	880	136
Checklist	27.8 (8.5-47.2)	14.3 (0.6-28.1)	91.5 (18.9-100.0)	39.3 (21.2-57.4)	73.5 (68.0-79.0)	276	28
Survey	16.5 (3.6-29.5)	33.0 (3.8-62.1)	89.6 (0.0-100.0)	38.6 (24.2-53.0)	87.1 (83.1-91.1)	312	44
Draw+Surv	16.5 (3.6-29.5)	5.8 (0.0-13.7)	96.9 (41.9-100.0)	27.3 (1.0-53.6)	83.9 (79.7-88.0)	312	11
Draw+List	22.4 (12.8-32.1)	20.0 (8.7-31.2)	97.2 (31.6-100.0)	67.6 (51.9-83.4)	80.8 (77.2-84.3)	514	34
Draw+Surv+List	18.0 (3.8-32.3)	8.5 (0.0-19.7)	98.1 (32.4-100.0)	50.0 (10.0-90.0)	83.0 (77.7-88.3)	197	6
Surv+List	18.0 (3.8-32.3)	16.9 (0.0-34.9)	95.7 (12.6-100.0)	46.2 (19.1-73.3)	84.0 (78.7-89.3)	197	13
Draw—Surv	20.8 (12.5-29.2)	32.7 (17.6-47.9)	84.3 (8.1-100.0)	35.5 (28.3-42.7)	82.7 (79.9-85.4)	880	169
Draw—List	20.8 (12.5-29.2)	32.2 (17.3-47.1)	83.5 (7.8-100.0)	33.9 (26.9-40.9)	82.4 (79.6-85.2)	880	174
Surv—List	21.8 (8.9-34.6)	25.9 (7.5-44.3)	88.3 (4.2-100.0)	38.0 (26.7-49.3)	81.1 (77.3-84.9)	479	71

Source: Authors' calculations. **Note:** Prevalence, sensitivity, and specificity include 95% confidence intervals derived as shown in Annex 6. Precision (PPV) represents the percentage of red-flagged cases confirmed as negligence through comprehensive psychological assessments.

Table A.8: *Results for Physical Maltreatment: Prevalence, Confusion Matrix Metrics, and Sample Size by Tool*

	Prevalence	Confusion Matrix				Observations	
Tool		Sensitivity	Specificity	PPV	NPV	Total	With red flag
Any tool	20.5 (12.1-28.8)	36.8 (19.7-53.9)	81.3 (0.4-100.0)	33.7 (27.1-40.3)	83.3 (80.5-86.1)	876	196
Drawing	20.5 (12.1-28.8)	27.9 (14.4-41.4)	87.1 (15.1-100.0)	35.7 (27.8-43.7)	82.4 (79.7-85.2)	876	140
Checklist	20.6 (2.3-38.8)	7.8 (0.0-18.2)	91.5 (28.2-100.0)	19.0 (2.3-35.8)	79.3 (74.1-84.5)	250	21
Survey	10.7 (0.4-21.0)	44.7 (0.0-89.7)	89.6 (0.0-100.0)	34.1 (19.6-48.7)	93.1 (90.0-96.2)	292	41
Draw+Surv	10.7 (0.4-21.0)	6.4 (0.0-16.8)	96.9 (33.7-100.0)	20.0 (0.0-44.8)	89.6 (86.1-93.2)	292	10
Draw+List	20.1 (11.2-29.0)	18.0 (7.2-28.7)	97.2 (33.1-100.0)	62.1 (44.4-79.7)	82.5 (79.1-85.9)	499	29
Draw+Surv+List	8.5 (3.1-13.9)	6.7 (0.0-19.3)	98.1 (28.5-100.0)	25.0 (0.0-67.4)	91.9 (87.8-96.0)	176	4
Surv+List	8.5 (3.1-13.9)	20.0 (0.0-40.2)	95.7 (7.5-100.0)	30.0 (1.6-58.4)	92.8 (88.9-96.7)	176	10
Draw—Surv	20.5 (12.1-28.8)	34.6 (18.4-50.8)	84.3 (5.5-100.0)	36.3 (29.1-43.5)	83.4 (80.6-86.1)	876	171
Draw—List	20.5 (12.1-28.8)	31.2 (16.4-46.1)	83.5 (8.0-100.0)	32.7 (25.7-39.8)	82.5 (79.7-85.3)	876	171
Surv—List	16.8 (4.7-29.0)	23.7 (3.5-43.9)	88.3 (0.1-100.0)	29.0 (17.7-40.3)	85.1 (81.6-88.6)	451	62

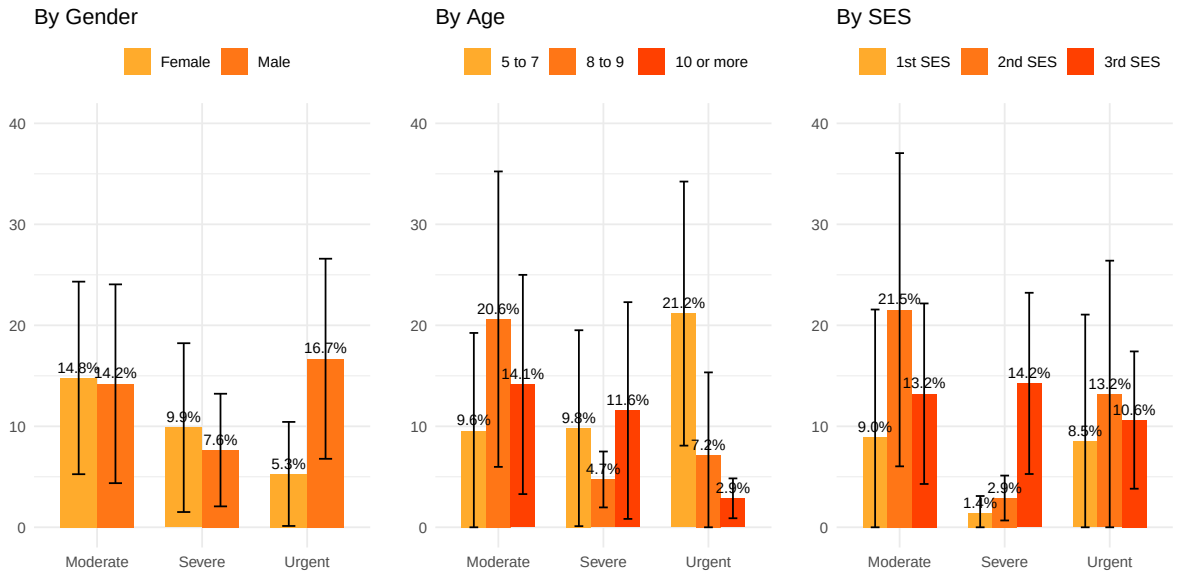
Source: Authors' calculations. **Note:** Prevalence, sensitivity, and specificity include 95% confidence intervals derived as shown in Annex 6. Precision (PPV) represents the percentage of red-flagged cases confirmed as physical maltreatment through comprehensive psychological assessments.

Table A.9: *Results for Emotional Maltreatment: Prevalence, Confusion Matrix Metrics, and Sample Size by Tool*

	Prevalence	Confusion Matrix				Observations	
Tool		Sensitivity	Specificity	PPV	NPV	Total	With red flag
Any tool	18.8 (10.7-26.9)	38.5 (20.0-57.0)	81.3 (0.0-100.0)	32.3 (25.7-38.9)	85.1 (82.4-87.8)	857	192
Drawing	18.8 (10.7-26.9)	25.4 (12.4-38.5)	87.1 (16.3-100.0)	31.3 (23.4-39.2)	83.5 (80.8-86.2)	857	131
Checklist	21.8 (3.8-39.8)	10.8 (0.0-23.4)	91.5 (21.9-100.0)	26.1 (8.1-44.0)	78.6 (73.3-83.9)	254	23
Survey	20.9 (6.5-35.3)	26.1 (4.4-47.8)	89.6 (0.0-100.0)	40.0 (25.7-54.3)	82.1 (77.7-86.6)	330	45
Draw+Surv	20.9 (6.5-35.3)	4.4 (0.0-10.1)	96.9 (49.8-100.0)	27.3 (1.0-53.6)	79.3 (74.9-83.8)	330	11
Draw+List	17.8 (9.8-25.7)	15.1 (5.4-24.8)	97.2 (36.2-100.0)	54.2 (34.2-74.1)	84.1 (80.8-87.5)	485	24
Draw+Surv+List	18.8 (4.6-33.0)	2.7 (0.0-8.2)	98.1 (52.1-100.0)	25.0 (0.0-67.4)	81.3 (75.8-86.8)	199	4
Surv+List	18.8 (4.6-33.0)	10.7 (0.0-23.3)	95.7 (26.1-100.0)	36.4 (7.9-64.8)	82.2 (76.7-87.7)	199	11
Draw—Surv	18.8 (10.7-26.9)	34.8 (17.8-51.7)	84.3 (3.7-100.0)	33.9 (26.7-41.2)	84.8 (82.1-87.5)	857	165
Draw—List	18.8 (10.7-26.9)	31.0 (15.6-46.4)	83.5 (6.6-100.0)	30.3 (23.3-37.3)	84.0 (81.2-86.7)	857	165
Surv—List	21.3 (8.4-34.2)	22.7 (5.8-39.6)	88.3 (7.7-100.0)	34.3 (23.0-45.7)	80.9 (77.1-84.7)	476	67

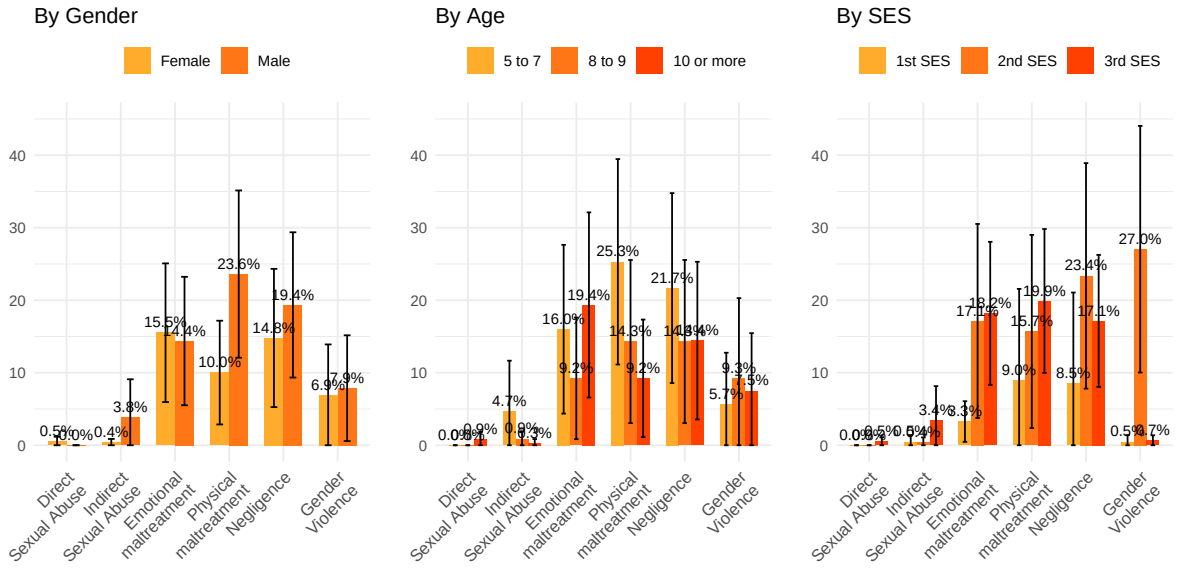
Source: Authors' calculations. **Note:** Prevalence, sensitivity, and specificity include 95% confidence intervals derived as shown in Annex 6. Precision (PPV) represents the percentage of red-flagged cases confirmed as emotional maltreatment through comprehensive psychological assessments.

Figure A.3: *Prevalence by Gender, Age, and Socioeconomic Status, by Severity*



Source: Authors' calculations. **Note:** Maltreatment cases are determined through comprehensive psychological assessments. Bars represent the percentage of children with confirmed maltreatment by severity. Whiskers indicate 95% confidence intervals.

Figure A.4: *Prevalence by Gender, Age, and Socioeconomic Status, by Type of Maltreatment*



Source: Authors' calculations. **Note:** Maltreatment cases are determined through comprehensive psychological assessments, except in cases previously identified by the school. Bars represent the percentage of children with confirmed maltreatment by type; whiskers indicate 95% confidence intervals. Approximate percentages (%) by subgroup: *Gender (female/male)* – Direct Sexual Abuse (0.5/0.0), Indirect Sexual Abuse (0.4/3.8), Emotional Maltreatment (15.5/14.4), Physical Maltreatment (10.0/23.6), Negligence (14.8/19.4), Gender Violence (6.9/7.9); *Age (5–7 / 8–9 / 10+)* – 0.0/0.0/0.9, 4.7/0.9/0.3, 16.0/9.2/19.4, 25.3/14.3/9.2, 21.7/14.3/14.4, 5.7/9.3/7.5; *School-area SES (1/2/3)* – 0.0/0.0/0.5, 0.5/0.4/3.4, 3.3/17.1/18.2, 9.0/15.7/19.9, 8.5/23.4/17.1, 0.5/27.0/0.7. SES refers to the school's surrounding area, not the individual child or family.

Annex 8. Cross-Tabulation of Index vs. Reference Results: Precision and Sensitivity for Different Index Test Thresholds

In this section, we present the cross-tabulation of the diagnostic tools scores (index tests) and the comprehensive psychological assessment (reference standard). We also present the plots that illustrate the relationship between varying thresholds for our diagnostic tools and their corresponding precision and sensitivity metrics. These plots provide valuable insights into the performance of our maltreatment detection tools across different threshold settings. By visualizing how precision and sensitivity change as we adjust the criteria for triggering a red flag, we can identify optimal threshold values that balance the trade-off between accurately identifying true cases of maltreatment (sensitivity) and minimizing false positives (precision).

We analyze three primary diagnostic tools: Teacher's Checklist, Piers-Harris Self-Concept Scale (Survey), and Projective Drawings. For each tool, we provide the cross-tabulation of its results against the outcome of the comprehensive psychological assessment and plot the precision and sensitivity against their respective thresholds. These visualizations allow us to observe trends, identify potential optimal operating points, and compare the performance characteristics of each tool.

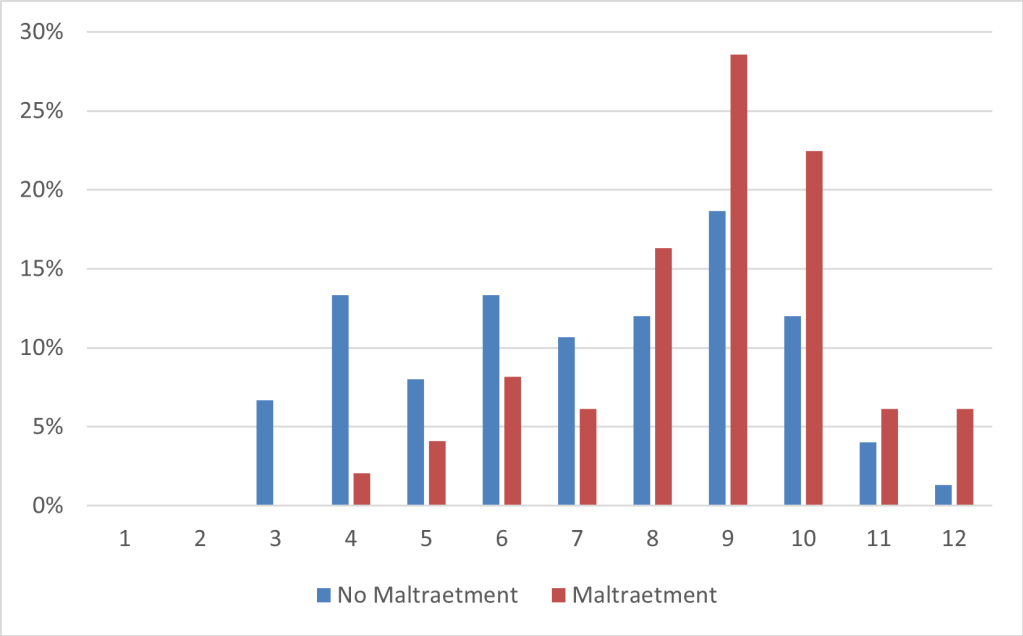
Piers-Harris Self-Concept Scale (Survey)

Table A.10: *Piers-Harris Self-Concept Scale Score Distribution (Positive vs. Negative Reference Standard): Frequency Table*

	No Maltreatment		Maltreatment	
Score Bin	# (Frequency)	Relative Frequency	# (Frequency)	Relative Frequency
0–5	0	0%	0	0%
5–10	0	0%	0	0%
10–15	5	7%	1	2%
15–20	10	13%	1	2%
20–25	6	8%	2	4%
25–30	10	13%	4	8%
30–35	8	11%	3	6%
35–40	9	12%	8	16%
40–45	14	19%	14	29%
45–50	9	12%	11	22%
50–55	3	4%	3	6%
55–60	1	1%	3	6%

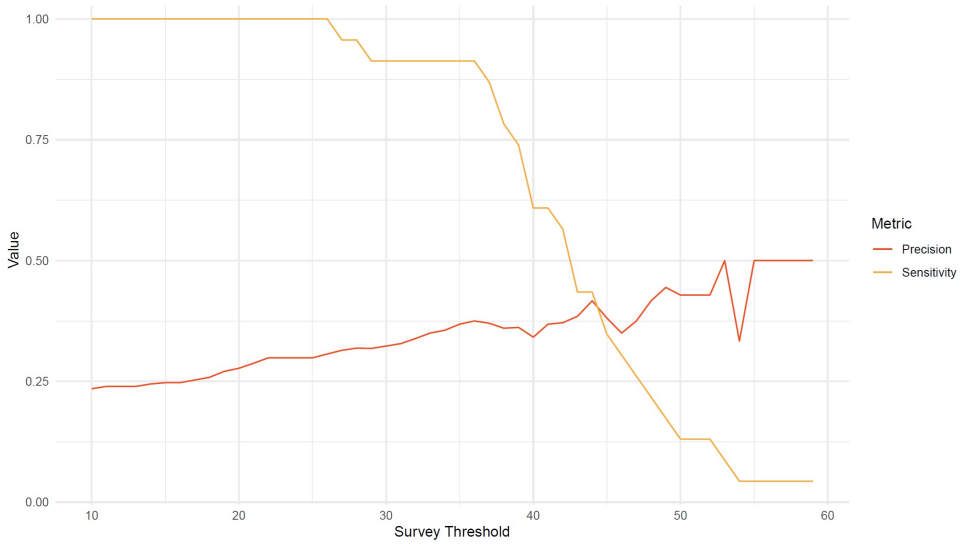
Source: Authors' calculations. **Note:** The table reports the distribution of Piers–Harris Self-Concept Scale total scores (0–60) in 5-point bins, separately for children with a negative ("No Maltreatment") versus positive ("Maltreatment") reference standard. "#" is the count in each bin. "Relative Frequency" is the within-group percentage. Percentages sum to 100% within each group (minor discrepancies may occur due to rounding).

Figure A.5: *Piers-Harris Self-Concept Scale Score Distribution (Positive vs. Negative Reference Standard): Histogram*



Source: Authors' calculations. **Note:** Histogram displays the distribution of Piers–Harris Self-Concept Scale total scores in 5-point bins, separately for children classified by the reference standard as No Maltreatment (blue) and Maltreatment (red). Bar heights are within-group percentages, so values sum to 100% within each group. Bin width = 5 score points (0–5, 5–10, . . . , 55–60). Minor discrepancies may occur due to rounding.

Figure A.6: *Precision and Sensitivity for Different Piers-Harris Self-Concept Scale Thresholds*



Source: Authors' calculations. **Note:** The figure shows how precision (positive predictive value) and sensitivity vary according to different thresholds applied to the Survey score for flagging cases.

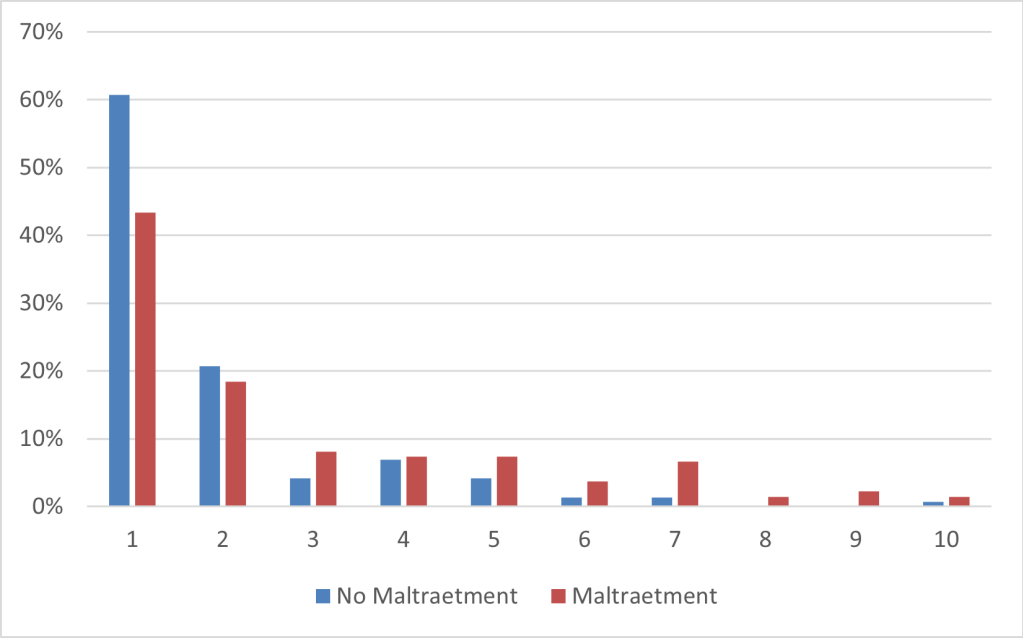
Teacher's Checklist

Table A.11: *Teacher's Checklist Score Distribution (Positive vs. Negative Reference Standard): Frequency Table*

	No Maltreatment		Maltreatment	
Score Bin	# (Frequency)	Relative Frequency	# (Frequency)	Relative Frequency
0–1	88	61%	59	43%
1–2	30	21%	25	18%
2–3	6	4%	11	8%
3–4	10	7%	10	7%
4–5	6	4%	10	7%
5–6	2	1%	5	4%
6–7	2	1%	9	7%
7–8	0	0%	2	1%
8–9	0	0%	3	2%
9–14	1	1%	2	1%

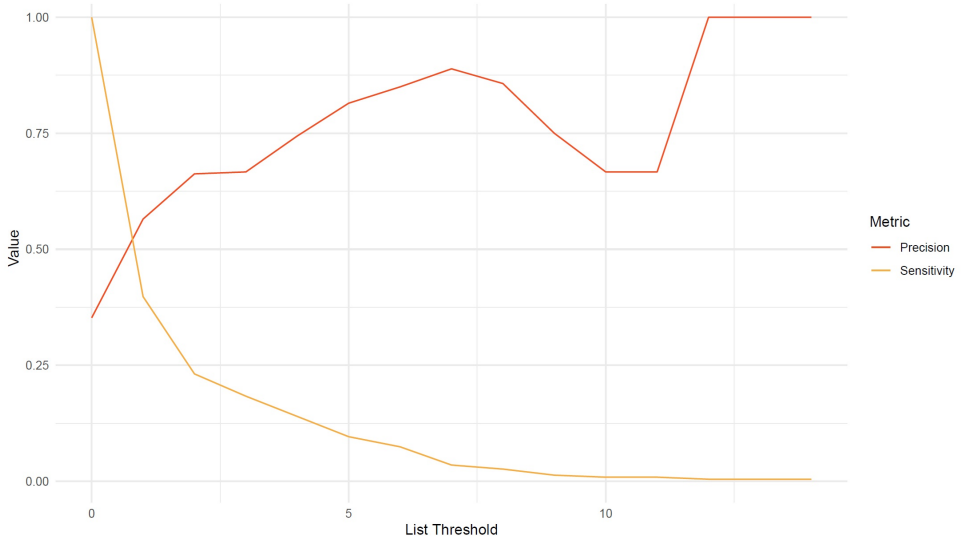
Source: Authors' calculations. **Note:** The table reports the distribution of the Teacher's Checlikst scores (i.e., number of selected items), separately for children with a negative ("No Maltreatment") versus positive ("Maltreatment") reference standard. "#" is the count in each bin; "Relative Frequency" is the within-group percentage. Percentages sum to 100% within each group (minor discrepancies may occur due to rounding).

Figure A.7: *Teacher's Checklist Score Distribution (Positive vs. Negative Reference Standard): Histogram*



Source: Authors' calculations. **Note:** Histogram displays the distribution of the Teacher's Checklist scores, separately for children classified by the reference standard as No Maltreatment (blue) and Maltreatment (red). Bar heights are within-group percentages, so values sum to 100% within each group.

Figure A.8: *Precision and Sensitivity for Different Teacher's Checklist Thresholds*



Source: Authors' calculations. **Note:** The figure shows how precision (positive predictive value) and sensitivity vary according to different thresholds applied to the checklist score for flagging cases.

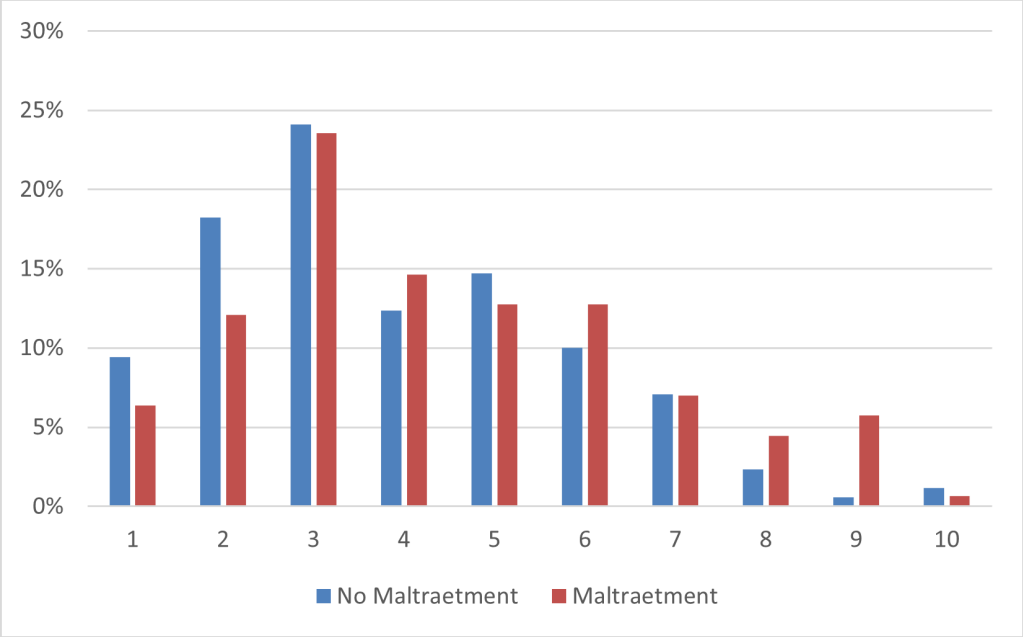
Projective Drawings

Table A.12: *Projective Drawings Score Distribution (Positive vs. Negative Reference Standard): Frequency Table*

	No Maltreatment		Maltreatment	
Score Bin	# (Frequency)	Relative Frequency	# (Frequency)	Relative Frequency
0–0.05	16	9%	10	6%
0.05–0.1	31	18%	19	12%
0.1–0.15	41	24%	37	24%
0.15–0.2	21	12%	23	15%
0.2–0.25	25	15%	20	13%
0.25–0.3	17	10%	20	13%
0.3–0.35	12	7%	11	7%
0.35–0.4	4	2%	7	4%
0.4–0.45	1	1%	9	6%
0.45–0.5	2	1%	1	1%

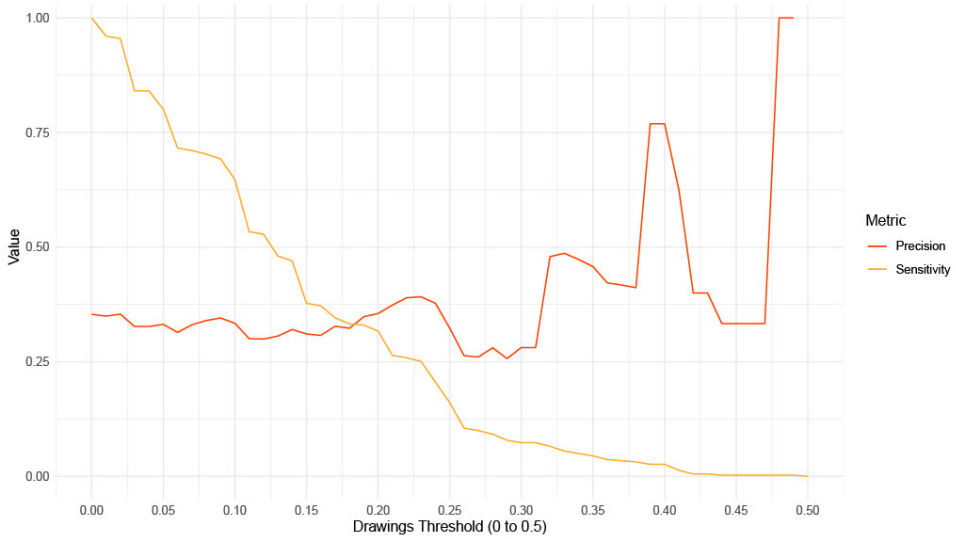
Source: Authors' calculations. **Note:** The table reports the distribution of the Projective Drawings scores (i.e., % of drawing indicators), separately for children with a negative ("No Maltreatment") versus positive ("Maltreatment") reference standard. "#" is the count in each bin; "Relative Frequency" is the within-group percentage. Percentages sum to 100% within each group (minor discrepancies may occur due to rounding).

Figure A.9: *Projective Drawings Score Distribution (Positive vs. Negative Reference Standard): Histogram*



Source: Authors' calculations. **Note:** Histogram displays the distribution of the Projective Drawings scores, separately for children classified by the reference standard as No Maltreatment (blue) and Maltreatment (red). Bar heights are within-group percentages, so values sum to 100% within each group.

Figure A.10: *Precision and Sensitivity for Different Projective Drawings Thresholds*



Source: Authors' calculations. **Note:** The figure shows how precision (positive predictive value) and sensitivity vary according to different thresholds applied to the Drawings score for flagging cases. Note that this score was not directly used to classify drawings.