Open Access

OBM Integrative and Complementary Medicine

LIDSEN Publishing Inc.

Research Article

Evaluating Screening Guidelines for Disruptive Behavior Problems in Children: A Systematic Review of the Accuracy of Parents' Concerns

Sarah L. Wells, Raymond H. Baillargeon *

Faculty of Health Science, University of Ottawa, Ottawa, Canada; E-Mails: swell044@uottawa.ca; raymond.baillargeon@uottawa.ca

* **Correspondence:** Raymond H. Baillargeon; E-Mail: raymond.baillargeon@uottawa.ca

Academic Editor: Marianna Mazza

Special Issue: Early Identification, Prevention and Care of Disruptive Behavior Problems in Very Young Children

OBM Integrative and Complementary Medicine
2025, volume 10, issue 1
doi:10.21926/obm.icm.2501009

Received: May 31, 2024
Accepted: December 15, 2024
Published: February 08, 2025

Abstract

Disruptive behavior problems (DBPs) in young children are early indicators of potential disruptive behavior disorders (DBD), which can lead to negative health and social outcomes. Secondary prevention strategies that target DBPs may facilitate early interventions and reduce these risks. Current Canadian pediatric practice guidelines provide an example one such strategy and suggest screening for DBPs only if a child's parent reports concerns about their behavior. This systematic review sought to determine if parents' concerns can provide enough information to justify a decision in favour of, or against, screening for DBPs. The protocol was registered on Prospero (CRD42021157492), and no funding was received. Six databases were searched (March 23–26, 2022) for prospective, retrospective, or naturalistic studies assessing the diagnostic accuracy of parents' concerns. Studies were included if they elicited parents' concerns about their child's behavior via an index test, used a reference standard to identify DBPs in children aged 0-5, and reported true/false positive and true/false negative outcomes. Studies were excluded if they did not include children in the target age range, did not report the outcomes of interest, used inappropriate sampling methods, measured heterogeneous mental health problems, elicited heterogeneous concerns from

parents, or if they were not a primary analysis of data. Risk of Bias was assessed using the QUADAS-2 tool, and results were synthesized to produce calibrated estimates of the accuracy of parents' concerns in the form of weighted kappa coefficients. Of 53 studies reviewed, only one met the eligibility criteria. Moderate agreement was found between the absence of DBPs and parents' concerns (k = 0.533, 95% CI: 0.501-0.564) and fair agreement for the presence of DBPs and parents' concerns (k = 0.255, 95% CI: 0.238-0.272). These findings suggest that parents' concerns alone may not be sufficiently accurate to guide clinical screening decisions, highlighting a significant gap in the literature. Further research is needed to validate this approach. Until more data becomes available, clinicians should be cautious when interpreting the presence or absence of parents' concerns about their child's behavior, and in using parents' concerns when making decisions to screen for DBPs.

Keywords
Systematic review; disruptive behaviors; conduct disorder; oppositional defiant disorder; diagnostic screening; diagnostic test accuracy; developmental surveillance

1. Introduction

1.1 An Overview of Disruptive Behavior Problems and Disorders

Disruptive behaviour problems (DBPs) can be conceptualized as a multidimensional framework of behaviours that evolve with psychological development across the life course [1]. Carter et al. identify non-compliance, temper-loss, low concern for others, and aggression as four core dimensions of DBPs, present throughout life [1]. The manifestation of DBPs shifts with developmental stages [1]; for example, a reflexive "no" to caregiver requests, and difficulty following directions from work supervisors both reflect non-compliant behaviour at different stages in an individual's development [1].

The presence of DBPs within this framework are associated with symptoms of disruptive behavior disorders (DBDs) [1], which are defined by the Diagnostic and Statistical Manual of Mental Disorders, 5th ed. (DSM-5) as "conditions involving problems in the self-control of emotions and behaviors" which manifest as "behaviors that violate the rights of others or conflict with societal norms or authority figures" [2]. Notable examples of DBDs include Oppositional Defiant Disorder (ODD) and conduct disorder (CD) as DBDs, both of which typically emerge as early as the preschool years [2]. Early signs of DBPs in infants and young children may indicate an elevated risk for the later emergence of DBDs [3, 4], presenting an opportunity for early prevention and intervention.

1.2 "The Burden of Suffering": The Impact of DBDs at an Individual and Population Level

There is consensus in the literature that healthcare strategies aimed at the prevention of childhood mental illness may greatly reduce the "burden of suffering" across the life course [5-8]. Offord et al. [8] define "burden of suffering" as the cumulative, holistic impact of the morbidity, continuity, prevalence, and the financial and social costs associated with a condition. Early

prevention strategies that target DBPs and DBDs may be especially important and can offer an opportunity to minimize the "burden of suffering" related to these disorders.

In the context of DBDs, the "burden of suffering" can be quite significant. A recent systematic review estimated that mental health disorders affect approximately 12.7% of the population aged 4-18 years in high income countries [9]. Within this pooled sample, the prevalence of ODD was estimated to be 3.3%, and conduct disorder was 1.3% [9]. Although prevalence estimates for younger children are less clear, mental health disorders, including DBPs, are known to affect around 18.4% of children aged 2 to 5 years [5, 10]. Further, morbidity related to the presence of DBDs in childhood can be significant [8]. Poor outcomes, such as delinquency, academic underachievement, and poor physical and mental health are common among children with DBDs [11], and the impacts of these outcomes may persist across the life course. For example, children diagnosed with CD have higher odds of experiencing depression, anxiety, substance use, and engaging in delinquent behaviors relative to their peers, which can lead to socioeconomic disadvantage and poor overall health outcomes in adulthood, and early mortality [12, 13]. In addition to the emotional, physical, and social costs, DBDs can also carry a significant financial burden at both the personal and public level due to increased public health care service use and the increased need for long-term psychological healthcare [14-18].

1.3 Minimizing the Burden of Suffering: Prevention of DBDs and DBPs

1.3.1 Secondary Prevention Strategies

Broadly, disease prevention strategies aim to reduce the burden of disease by targeting different stages of disease/illness development [19]. Given that the "burden of suffering" associated with DBPs and later DBDs can be significant, preventive strategies that target early DBPs may present an important opportunity to minimize this burden [20, 21]. In this context, secondary prevention strategies are therefore of particular interest, which focus on early identification of at-risk children before the onset of more severe behavioral disorders – thus introducing the opportunity for early intervention following early identification. For example, secondary prevention strategies that incorporate early screening during Well-Baby visits have been proposed as one strategy by which DBPs and DBDs can be detected [22, 23].

1.3.2 Universal and Targeted Screening

Screening plays a crucial role in secondary prevention strategies – in this context, by detecting early signs of DBPs. Targeted screening may focus on children deemed at elevated risk due to factors like family history or early behavioral indicators. In contrast, universal screening applies to all children within a certain age range, regardless of risk. Both approaches aim to prevent the progression of DBPs to (more severe) DBDs by facilitating early detection and intervention [22, 23]. Some research has suggested that universal screening may be particularly beneficial in detecting DBPs, given that early symptoms often emerge in preschool years and tend to persist as children age [24, 25]. However, the resources needed to implement a universal screening program can be high relative to those associated with a targeted screening strategy. It is therefore of interest to determine if targeted screening strategies exist that are effective in identifying DBPs in young children.

1.3.3 Current Recommendations

One systematic review of randomized clinical trials of disruptive behavior interventions found that secondary prevention strategies are most effective at preventing DBPs in preschool-aged children [21]. Several targeted intervention programs—such as the Family Checkup, Triple P, and Incredible Years Group Parenting programs—have shown success in reducing negative outcomes in at-risk children [21]. However, routine early developmental monitoring during Well-Baby visits, which could detect early signs of DBPs, was conspicuously absent from the interventions studied. Such early monitoring programs may represent an excellent opportunity for early detection and prevention of more severe behavioral disorders.

Within this context, it is significant that the Canadian Paediatric Society has published a clinical practice guideline that uses early screening for DBPs as part of a targeted, secondary prevention strategy [26]. These guidelines recommend that parents' concerns be used to inform screening decisions for DBPs in children beginning at the 24-month well-child visit and continuing until the child reaches 5 years of age. It advocates for primary care providers (PCPs) to routinely elicit parents' concerns to assist in deciding in favour of/against screening for DBPs, and that *if* parents' concerns are present, then PCPs should always decide in favour of screening. Otherwise, when parents' concerns are absent, PCPs should decide against screening [26]. Crucially, a summary analysis of the accuracy of parents' concerns in this context would provide a solid basis of evidence upon which we can inform the creation of similar targeted screening strategies. In this systematic review, we sought to fill this gap in the literature, and to determine what, if any, evidence exists that would support the use of parents' concerns in the context of screening decisions for DBPs.

1.4 Research Objective and Questions

1.4.1 Research Objective

The purpose of this research was to determine, via a systematic review, whether parents' concerns can provide PCPs with enough information to decide in favour of/against screening for DBPs.

1.4.2 Research Questions

First, we sought to determine if the mere presence of parents' concerns justifies deciding in favour of screening for DBPs in children. Second, we sought to determine if the mere absence of parents' concerns justifies deciding against screening for DBPs in children.

2. Methods

2.1 PITO Elements & Eligibility Criteria

PITO is a conceptual framework that underpins the construction of a systematic review of diagnostic test accuracy studies, identifying *a priori* the population of interest (P), the index test (I), the target condition (T), and the outcomes of interest (O) in eligible studies. The population of interest in this review were children between the ages of 0-5 years, and their parent/caregiver. The index test was the presence/absence of parents' psychological concerns about their child's

behaviour. Herein, psychological concerns are defined as any concern about the social, emotional, and/or behavioral development of the child. The reference standard was the presence or absence of DBPs in our population of interest.

The outcomes of interest that were assessed by this study are as follows:

1. The proportion of children with a DBP and whose parents have behavioural concerns (true positives, or TP)
2. The proportion of children with a DBP whose parents do not have behavioural concerns (false negatives, or FN)
3. The proportion of children without a DBP and whose parents do not have behavioural concerns (true negatives, or TN)
4. The proportion of children without a DBP and whose parents have behavioural concerns (false positives, or FP)

Studies that met the following eligibility criteria were included for review:

1. had investigated DBPs in children between the ages of 0-5 years of age
2. had systematically elicited parents' psychological concerns about their child via a validated index test (e.g., the Parental Evaluation of Developmental Status (PEDS) [27] or equivalent), or via a custom index test designed to elicit parents' concerns
3. the study confirmed the presence/absence of a DBP in children via a reference standard that had been validated for use within the population of interest (e.g., Eyberg Child Behavior Inventory (ECBI) [28], or equivalent)
4. the study reported the 4 outcomes of interest identified by this review or reported outcome measures from which the 4 outcomes of interest could be generated.

We excluded studies that met one or more of the following exclusion criteria:

1. the study population of interest did not match the review's population of interest (i.e., sampled only children 5 years and older, or did not report data for the population of interest that was separable from older populations)
2. the study reported the wrong outcomes (i.e., reported the wrong outcomes of interest, or did not report outcomes from which the 4 outcomes of this review could be generated)
3. the study measured heterogeneous mental health problems (i.e., measured a combination of problems from different developmental domains (e.g., DBPs and physical health problems)), or did not measure DBPs at all
4. the study elicited heterogeneous concerns from parents (e.g., concerns about psychological and physical development in their child), or elicited concerns from parents that were not psychological in nature (i.e., not about social, emotional, or behavioral development)
5. the study used a pseudo-retrospective or pseudo-prospective sampling method [29] (see below)
6. the study was not a primary analysis of data (e.g., re-analyzed data from a previous study)

There are two primary arguments justifying these eligibility/exclusion criteria. These can be categorized as 1) Sampling Methods, and 2) Defining Parents' Concerns.

2.1.1 Sampling Methods

Among diagnostic test accuracy (DTA) studies that have assessed the accuracy of parents' concerns with respect to DBPs, there are three sampling methods by which participant data can be

generated: naturalistic, prospective, and retrospective sampling. A description of each of these three sampling methods, according to the methods of Kraemer [29] are provided below.

Naturalistic sampling simply involves drawing a representative sample from the population of interest (N0) and testing every participant with both the reference standard and the index test irrespective of the participant's results on either test. From the results of these tests, a 2 × 2 contingency table can be populated and the 4 outcomes of interest (i.e., TP, FN, FP, TN) can be estimated. These outcomes are then used to estimate the values of P, Q, sensitivity (SE), specificity (SP), positive predictive value (PPV), and negative predictive value (NPV) (see *Data Synthesis and Analysis*). Briefly, P is the proportion of children who have a DBP within the population of interest, and Q is the proportion of parents' who have concerns within the population of interest. SE is the conditional probability of a parent having concerns if their child has a DBP; SP is the conditional probability of a parent *not* having concerns if their child does *not* have a DBP; PPV is the conditional probability of a child having a DBP if their parent has concerns; NPV is the conditional probability of a child *not* having a DBP if their parent does *not* have concerns.

Prospective DTA studies first draw a representative sample of parents from the population of interest and test all individuals in the sample with the index test (N0). From these index test results, we can estimate the proportion of parents in the population of interest with a positive index test result – that is, those who have concerns about their child's behaviour (Q). A second, random sample of children whose parents gave a positive response on the index test (N1) is then drawn from N0. A third, random sample of children whose parents gave a negative response on the index test (N2) is also drawn from N0. All children from the N1 and N2 samples are then administered the reference standard. From these reference standard results, we can estimate the conditional probability of a child having a positive reference standard result (i.e., DBP) if their parent had a positive index test result (i.e., concerns) – that is, we can estimate PPV. Similarly, we can estimate the conditional probability of a child having a negative reference standard result (i.e., no DBP) if their parent had a negative index test result (i.e., no concerns) – that is, we can estimate NPV. Having estimates of Q, PPV, and NPV, we can then proceed to estimate the 4 outcomes of interest, SE, SP, and P [29]. However, in the absence of the first sampling stage (N0) (i.e., "pseudo-prospective sampling" [29]), we cannot generate a prior estimate of Q. Without Q, we cannot calculate later estimates of the 4 outcomes of interest, SE, SP, and P [29], nor can we calibrate these measures (see *Data Analysis and Synthesis*) [30]. Therefore, pseudo-prospective DTA studies were unsuitable for this study's analysis protocol and were excluded from this review.

Inversely, retrospective DTA studies first draw a representative sample of children from the population of interest and test all individuals in the sample with the reference standard (N0). From these reference standard results, we can estimate the proportion of children in the population of interest with a positive reference standard result – that is, those who have a DBP (P). A second, random sample of children who received a positive result on the reference standard (N1) is then drawn from N0. A third, random sample of children who received a negative result on the reference standard (N2) is also drawn from N0. All parents of the children from the N1 and N2 samples are then administered the index test. From these index test results, we can then estimate the probability of a parent having a positive index test result (i.e., concerns) if their child has a positive reference standard result (i.e., a DBP) – that is, we can estimate SE. Similarly, we can estimate the conditional probability of a parent having a negative index test result (i.e., no concerns) if their child has a negative reference standard result (i.e., no DBP) – that is, we can estimate SP. Having

estimates of P, SE, and SP, we can then proceed to estimate the 4 outcomes of interest, PPV, NPV, and Q. However, in the absence of the first sampling stage (N0) (i.e., "pseudo-retrospective sampling" [29]), we cannot generate a prior estimate of P. Without P, we cannot calculate later estimates of the 4 outcomes of interest, NPV, PPV, and Q [29], nor can we calibrate these measures (*see Data Analysis and Synthesis*) [30]. Therefore, pseudo-retrospective DTA studies were unsuitable for this study's analysis protocol and were excluded from this review.

2.1.2 Defining Parents' Concerns

Determining which types of parents' concerns to consider can miss the target condition (i.e., DBPs), either by mapping too broadly (i.e., including too many types of concerns) or by mapping too narrowly (i.e., excluding too many types of concerns). There are limitations to either approach. For example, broadening the definition of behavioural concerns too much, such that many concerns that are included *do not* map onto DBPs (but, rather, on to other developmental problems), may artificially reduce our SP estimate. Alternatively, narrowing the definition of behavioural concerns too much, such that many concerns are excluded that *do* map onto DBPs may artificially reduce our SE estimate. Either situation will introduce bias. In effort to minimize the number of behavioral concerns that may be missed by defining this construct too narrowly, this review defined parents' concerns as concerns about the psychological development of their child. As mentioned, psychological development has been defined as the social, emotional, and/or behavioral development of the child. Therefore, studies that elicited concerns from parents that extended beyond the scope of this construct were excluded from this review.

2.2 Systematic Search Strategy

Eligible studies were identified using a systematic search strategy that was adapted for use in Medline(Ovid), Embase(Ovid), Central(Ovid), CINAHL(EBSCOHost), PsycINFO(Ovid), and Eric(Ovid) and Scopus (see Appendix A). Search results were first uploaded to Covidence Systematic Review Software [31], in which all steps of the eligibility review process were conducted. Following the import of search results, duplicates were removed. Title-and-Abstract Screening was then completed, where potentially relevant studies were selected for full-text review. Consensus on decisions to include or exclude was required between the 2 reviewers assessing each record, and conflicts were resolved via discussion.

Following Title-and-Abstract Screening, full-text articles were independently screened for eligibility by 2 reviewers, with three reviewers in total contributing to this stage of review (SW, RB & HP). Studies were excluded from our analysis if they did not meet our eligibility criteria (see below). Like the Title and Abstract screening phase, consensus was required between 2 reviewers on decisions to include or exclude a study. In situations where a reviewer wished to vote to exclude a record, the reviewer had to select one primary reason for ineligibility in Covidence. Any disagreement between reviewers was resolved via discussion between all 3 full-text reviewers. Consensus among 2/3 of the reviewers was necessary resolve disagreements regarding whether to include or exclude a study, and regarding the primary reason for exclusion for any given study.

Following Full-Text Review, backward citation searching was completed by manually checking the reference lists of the included studies [32]. One reviewer (SW) imported all reference material from included studies into Covidence for completion of screening. Title and Abstract screening, and

full-text screening during this phase were completed independently by two reviewers, and conflicts were resolved again via discussion. If a study(s) was missing data required for analysis, the review team contacted the study author(s) via email to request the missing data during this review phase.

2.3 Data Extraction

It is important to note here that the planned data extraction and analysis methods used by this review depended on the sampling methods used by each included study. That is, for example, the extraction and analysis of data would have been slightly different for a study with a prospective sampling design than for one with a retrospective sampling design. Similarly, the exact methods required to analyze data from a study with a naturalistic sampling design are also unique. Given the results of this review, all extraction and analysis methods described herein are specific to data from studies that used a naturalistic sampling design.

A data extraction form specific to naturalistic sampling methods was adapted from the "Data Collection form for intervention review – RCTs and non-RCTs" data collection template produced by the Cochrane Collaboration [33]. An example of the adapted data extraction form for a naturalistic sampling design can be found in Appendix B. Data pertaining to the outcomes of interest (i.e., TP, TN, FP, and FN), participant demographics (i.e., age, sex, gender, … etc.), publication details (i.e., authors, date of publication, … etc.), and study characteristics (i.e., number of participants, reference standard used, index test used etc.) were independently extracted from included studies and entered into the data extraction form.

2.4 Assessment of Risk of Bias

As per the PRISMA-DTA guidelines, risk of bias was assessed in included studies based on their applicability to the current problem of interest [34]. Applicability of studies to the current review and other sources of methodological bias were assessed using the QUADAS-2 tool [35]. According to the QUADAS-2, the applicability of a DTA study is defined as the "the extent to which primary studies are applicable to the review's research questions" and bias is defined as "systematic flaws or limitations in the design or conduct of a study [which] distort the results" [35].

For each included study, two reviewers (SW & RB) independently completed the QUADAS-2 assessment to determine the risk of bias within each of the following 4 domains: patient selection methods, the index test, the reference standard, and the Flow and Timing of test administration. Reviewers also used the QUADAS-2 tool to determine whether each included study was applicable to our review questions. This was assessed across each of the following 3 domains: the patient selection methods, the index test, and the reference standard. Following independent assessment, any discrepancies in reviewers' judgments for each of the domains were resolved via discussion. The QUADAS-2 poses signalling questions for each risk of bias domain, and each applicability domain [36] to assist reviewers in their assessment. Reviewers judged the risk of bias as "low", "unclear", or "high" [36] based on their responses to each signalling question. As per the QUADAS-2 guidance document, a response of "no" to any signalling question indicated the potential for bias; a response of "unclear" was appropriate when the included study reported insufficient data to permit a judgment from reviewers; a response of "yes" to any signalling question indicated a low risk of bias [36].

2.5 Data Analysis & Synthesis

2.5.1 Naturalistic Sampling Designs

To determine whether the mere presence of parents' concerns justifies deciding in favour of screening for DBPs in children, reviewers needed to determine how accurately parents' concerns could rule in the presence of a DBP in children. To do this, reviewers calculated the concerns' positive predictive value (PPV), specificity (SP), and jackknife estimates of RR1 based on the four outcomes of interest extracted from each included study. RR1 was defined as the ratio of the odds of having a DBP among those children who have parents with concerns, over the odds of having a DBP among all children [29]. Standard errors, and 95% confidence intervals were obtained for all measures listed above. The formulae used to calculate SP, PPV and RR1 for a naturalistic sampling design can be found in Figure 1.

$$SP = \frac{TN}{P'} \qquad SE = \frac{TP}{P}$$

$$PPV = \frac{TP}{Q} \qquad NPV = \frac{TN}{Q'}$$

$$RR1 = \frac{SE}{(1-SP)} \qquad RR3 = \frac{SP}{(1-SE)}$$

$$k(0,0) = \frac{(SP-Q')}{Q} \qquad k(0,0) = \frac{(PPV-P)}{P'} \qquad k(0,0) = \frac{(RR1-1)}{(RR1+(P'/P))}$$

$$k(1,0) = \frac{(SE-Q)}{Q'} \qquad k(1,0) = \frac{(NPV-P')}{P} \qquad k(1,0) = \frac{(RR3-1)}{(RR3+(P/P'))}$$

Figure 1 Necessary Formulae for Data Analysis for a Naturalistic Sampling Design. *Note. P* refers to the proportion of children with a DBP (*P = TP + FN*, and *P'* = (1 - P)). *Q* refers to the proportion of children of concerned parents (*Q = TP + FP*, and *Q'* = (1 - Q)). *TP* refers to the proportion of children who have a DBP and have parents with concerns; *FP* refers to the proportion of children without a DBP and have parents with concerns; *FN* refers to the proportion of children with a DBP and have parents without concerns*; TN* refers to the proportion of children without a DBP and have parents without concerns.

For each included study, these estimates were then calibrated on a common scale (ranging from 0 to 1) using a weighted kappa coefficient, k(0,0) [29]. The formulae used to calibrate SP, PPV and RR1 can also be found in Figure 1. Note that, once calibrated, all three estimates yielded the same jackknife estimate of k(0,0) [29]. One may interpret a weighted k(0,0) value as a quality index of the reproducibility of a positive diagnosis, where k(0,0) equal to 0.0 indicates that any agreement between a positive result on the reference standard and a positive result on the index test is due to chance alone [29]. Conversely, k(0,0) equal to 1.0 indicates 100% corrected-for-chance agreement between a positive result on the reference standard and on the index test. As defined by Landis & Koch [36], the relative strength of agreement represented by any weighted kappa statistic ranges from poor ($k \leq 0.00$) to almost perfect ($0.8 < k \leq 1.00$) (see Table 1). Values of k(0,0) must be at least

substantial (i.e., k > 0.6) to demonstrate that parents' concerns can provide a reliable means to rule *in* the presence of a DBP in the absence of additional information [37].

Table 1 Strength of Agreement Represented by Weighted Kappa Statistics.

Kappa statistic	Strength of Agreement
$k \leq 0.00$	Poor
$0.00 < k \leq 0.20$	Slight
$0.20 < k \leq 0.40$	Fair
$0.40 < k \leq 0.60$	Moderate
$0.60 < k \leq 0.80$	Substantial
$0.80 < k \leq 1.00$	Almost Perfect

Note. This table is adapted from the thresholds defined and published by Landis & Koch [37].

To determine if the mere absence of parents' concerns justifies deciding against screening for DBPs in children, reviewers needed to determine how accurately parents' concerns could rule out the presence of a DBP in children. This was done by calculating the concerns' negative predictive value (NPV), sensitivity (SE), and jackknife estimates of RR3 based on the four outcomes of interest extracted from each included study. RR3 was defined as the ratio of the odds of not having a DBP among children who have parents with no concerns, over the odds of not having a DBP among all children [29]. Standard errors and 95% confidence intervals were also obtained for all measures listed above. The formulae used to calculate SE, NPV and RR3 for a naturalistic sampling design can also be found in Figure 1.

These estimates were also calibrated on a common scale (ranging from 0 to 1) using a weighted kappa coefficient, k(1,0) [29]. Formulae for calibrating SE, NPV, and RR3 can be found in Figure 1. Note that, once calibrated, all three estimates yielded the same jackknife estimate of k(1,0) [29]. One may interpret a weighted k(1,0) value as a quality index of the reproducibility of a negative diagnosis, where values of k(1,0) equal to 0.0 indicate that any agreement between a negative result on the reference standard and a negative result on the index test is due to chance alone [29]. Conversely, values of k(1,0) equal to 1.0 indicate 100% corrected-for-chance agreement between a negative result on the reference standard and on the index test. Values of k(1,0) exceeding 0.6 indicate that parents' concerns can provide a reliable means to rule *out* the presence of a DBP [37].

2.6 Meta- and Subgroup Analyses

In the review protocol, a meta-analysis plan was proposed which used a bivariate, multilevel model of k(0,0) and k(1,0) estimates to construct a summary Receiver Operating Characteristic (sROC) curve [38]. The area under the sROC curve would have then been calculated to determine the overall accuracy of parents' concerns. However, one scenario in which the proposed bivariate model may fail to converge arises when too little data is available for meta-analysis (i.e., an insufficient number of studies is returned from our search that meet our eligibility criteria). There is little consensus on what constitutes "too little" data; however, by convention, attempting to use data from fewer than 4 studies will likely render the model unstable. Therefore, reviewers defined, *a priori*, that a minimum threshold of 4 included studies was required to complete a meta-analysis. As is discussed in Section 3, the final systematic search did not yield a sufficiently large dataset from

which summary estimates of k(1,0) and k(0,0) could be modeled. Therefore, the meta-analysis method published in the protocol was not used to meta-analyze data from the included study.

3. Data Analysis and Results

3.1 Study Selection

Figure 2 was adapted from the PRISMA flow diagram [39] and summarizes the results of the screening and review process. Systematic searches returned 8420 articles, and backward citation searching returned an additional 19 articles. 3475 articles were removed as duplicates. Titles and abstracts from the remaining 4964 records were each screened independently by 2 reviewers, with 6 reviewers in total contributing to this screening phase (SW, HP, TC, IM, NY & RB). 4910 records were considered irrelevant, and thus excluded at this stage. Of the remaining 54 articles, 1 full-text version [40] was unable to be retrieved. 53 full-text articles were independently screened for eligibility by 2 reviewers, with 3 reviewers in total contributing to this screening phase (SW, HP & RB). 52 out of 53 full-text records were ineligible for inclusion for a variety of reasons. Just one article [41] met all inclusion criteria. Very few studies failed to meet inclusion criteria for one reason alone. However, constraints within the Covidence software required reviewers to select only one, primary reason for exclusion on which reviewers could reach a consensus (see Figure 2).

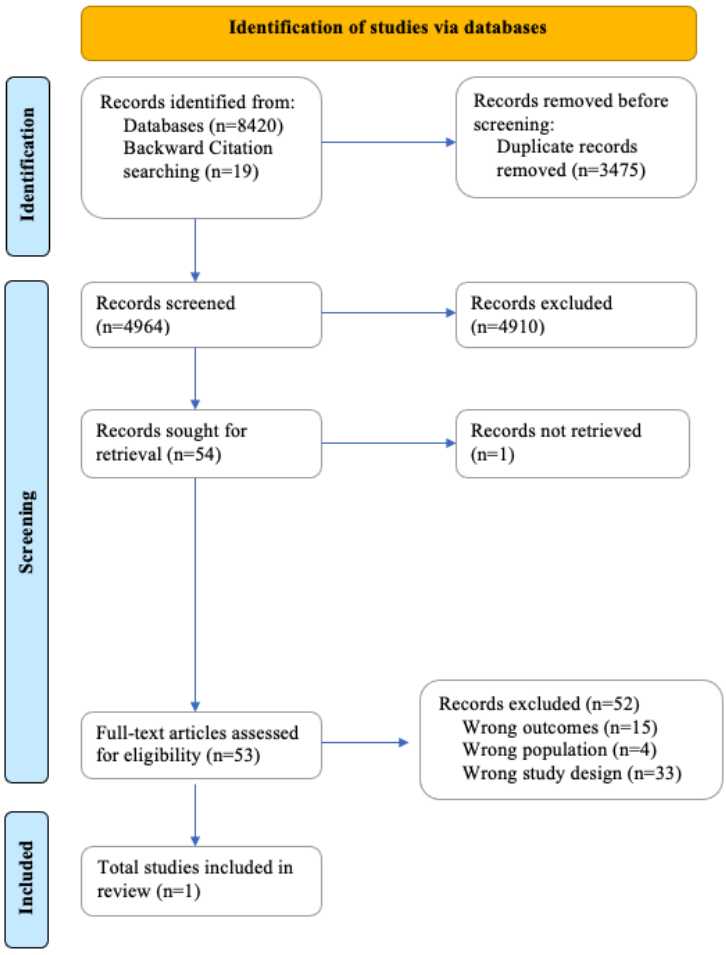


Figure 2 Prisma Flow Diagram. *Note.* This diagram was adapted from the *Prisma Flow Diagram* published by Page et al., 2021.

For each of the 52 records excluded at the full-text screening stage, Table 2 provides a description of how the selected reason to exclude a study given in Covidence aligned with our exclusion criteria. If a study was missing data that were required for analysis, the corresponding author(s) were contacted to request the missing data. One such study was retrieved via our search [42]. In this study, sample data published for participants under the age of 5 could not be analyzed separately from sample data corresponding to participants over the age of 5. As such, we could not generate necessary outcome measurements (i.e., TPs, FNs, TNs, and FPs) for our population of interest using only the data published by Reijneveld et al. [42]. A request for the disaggregated data was sent to the study's authors. Unfortunately, no response was received by the time of submission. Therefore, the paper by Reijneveld et al. [42] was excluded from the analysis.

Table 2 Description of Primary Reasons for Exclusion in Covidence.

Exclusion criterion met based on reason-for-exclusion given in Covidence	Description of primary reason-for-exclusion in Covidence	Excluded studies
Wrong Outcomes (*n* = 15)	Study did not report proportions of TPs, TNs, FPs, or FNs as outcome, or did not report similar outcomes that would allow us to generate estimates of proportions of TPs, TNs, FPs, and FNs (*n* = 15)	References [42-56]
Wrong Population of Interest (*n* = 4)	Results from children in the target population (i.e., under the age of 5 years) were not available or could not be analyzed (*n* = 4)	References [41, 57-59]
Wrong Study Design (*n* = 33)	Incomplete sampling methods (*n* = 7)	References [60-66]
	Re-analysis of existing data (*n* = 2)	References [67, 68]
	Elicited heterogeneous concerns from parents (*n* = 4)	References [69-72]
	Measured heterogeneous developmental problems (*n* = 6)	References [73-78]
	Absent or inappropriate index test (*n* = 7)	References [79-85]
	Absent or inappropriate reference standard (*n* = 7)	References [86-93]

Three other studies were identified in Covidence as having samples that did not match our population of interest and were therefore excluded from this review. First, the study by August et al. [58] included "children [that] ranged in age from 7-12 years", and thus contained a sample that was older than our target population. Similarly, the study by Poduska [59] sampled from "all first-graders in 9 public elementary schools … [that were] participants in 2-school based interventions targeting early learning and aggressive behavior". However, the study did not report the mean age or age range of children included in their sample. In North America, most children in first-grade are between 6 and 7 years of age. Hence, the sample studied by Poduska [59] appeared to be older than our population of interest and the study was therefore excluded from this review. Lastly, the study by Feldman et al. [60] was erroneously categorized in Covidence as containing the wrong population

of interest. However, the children included in the study by Feldman et al. [60] were all 2 years of age, and thus were within this review's population of interest. Nonetheless, all children included in this sample had been diagnosed with, or were at risk for, developmental delays [60]. Therefore, this study did not match this review's target condition (DBPs), and therefore the study was excluded on the grounds that the developmental problems measured were too heterogeneous.

3.2 Heterogeneity in Parents' Concerns

As previously mentioned, a relatively broad definition of parents' concerns was employed by this review, where studies that elicited parents' psychological concerns (i.e., social, emotional, or behavioral concerns) about their young children were eligible. Parents' concerns were considered *too* heterogeneous if they addressed psychological development and other domains of development in their child simultaneously. For example, a study that elicited concerns from parents that addressed physical or cognitive development were not eligible for inclusion. Four studies were identified during full text-screening that elicited parent concerns which were considered too heterogeneous for the scope of this review. Summaries of how parents' concerns were defined and measured by each of these 4 studies, and why each study was ultimately excluded from this review are presented below.

3.2.1 Barkley, Shelton, Crosswait, et al., 2002 [70]

In the study by Barkley et al., there were no measurement tools used that could be considered index tests that elicited parents' concerns. The only measure that came close to an index test (as defined by this review) was the Normative Adaptive Behavior Checklist (NABC). However, this tool did not appear to elicit specific concerns from parents about their child's psychological development. Further, the NABC assesses global developmental functioning across six very different domains: fine and gross-motor skill development, sensory-motor skills, language, self-help skills, independent living skills, and social skills [70]. Therefore, reviewers considered the data gathered by Barkley et al. via the NABC to be too heterogeneous for this review.

3.2.2 Karabekiroglu, Uslu, Kapci-Seyitoglu, et al., 2013 [71]

In the study by Karabekiroglu et al., there were again no measurement tools used that could be considered index tests that elicited parents' concerns. The only measure that appeared close to an index test was a series of questionnaire items that asked parents to "state the mental and/or developmental problems that [parents] thought to exist in [their] child" [71]. Importantly, the questionnaire items did not appear to elicit specific psychological concerns or worries from parents. Further, items on this questionnaire appeared to address areas of global developmental functioning (i.e., speech and language, learning/comprehension, sleeping, eating/feeding behavior) and areas of social-emotional development (i.e., irritability, phobias/fearfulness, aggression, hyperactivity, attention problems, excessive crying, and social withdrawal) [71]. Therefore, the "index test" used by Karabekiroglu et al. [71] was considered too heterogeneous for the scope of this review.

3.2.3 Kruizinga, Jansen, de Haan, et al., 2012 [72]

In the study by Kruizinga et al., there were again no measurement tools used that could be considered index tests that elicited parents' concerns. The only measure that came close to an index test (as defined by this review) was an item which assessed whether parents "worried about their child's upbringing" [72]. It was unclear to reviewers whether this measure was contained within a validated screening tool, or if it was a standalone item. Nonetheless, reviewers felt that "upbringing" as a construct was too broad and did not fit this review's definition of parents' concerns.

3.2.4 Wendland, Danet, Gacoin, et al., 2014 [73]

The study by Wendland et al. sought to examine the psychometric properties of a French translation of the BITSEA, which is a screening tool for the identification of elevated levels of social-emotional behavior problems or delayed levels of competence in children aged 1-3 years [73]. The BITSEA contains 2 single-item questions that assess parents' worries, in addition to subscales that measure behavioral problems. Pertinent to the scope of this review, the BITSEA-A item asks parents to report on the level of worry they have about their young child's behavior, emotions, or relationships [73]. In their analysis, Wendland et al. calculated correlations between BITSEA subscale scores and the level of parental worry as reported by parent responses to the BITSEA-A item [73]. Unfortunately, this study was erroneously categorized during the full-text screening phase as having reported heterogeneous concerns. Given that the BITSEA-A screening item does elicit specific concerns from parents, the study did in fact meet this inclusion criterion. Nonetheless, the fact that only correlational data was reported by Wendland et al. (from which we could not generate estimates of proportions of TPs, TNs, FPs, and FNs) rendered this study ineligible for inclusion.

3.3 Included Study Characteristics: Glascoe, MacLean & Stone, 1991 [41]

3.3.1 Demographic and Study Characteristics

Demographic and study characteristic data for the included study by Glascoe et al. [41] was extracted by reviewers (SW & RB). The study's stated purpose This cross-sectional study enrolled 99 parent-child dyads from five sampling sites, where patients who had visited one of the five sampling sites to seek non-acute medical care were eligible to participate in the study [41]. Sampling sites consisted of "two urban teaching hospitals" and "three private pediatric practices" [41]. Four eligible parent/child dyads declined participation, leading to a sample size of n = 95. Children enrolled were between 24-78 months of age, with a mean age of 48 months [41]. Statements that disclosed the study's funding were not included in the publication by Glascoe et al. [41].

3.3.2 Eliciting Parents' Concerns: The Index Test

First, an index test was administered verbally to parents [41]. The first question on the index test "asked each parent 'Please tell me any concerns about your child's learning and development?'" [41]. This generated positive responses specific to behavioral concerns from 8 out of 95 parents. All parents were then asked, "Do you have any concerns about the way [your child] behaves" [41]. This second question generated 26 positive responses [41]. Results from these 2 index test questions

were pooled to yield a total of 34 parents that had indicated that they had concerns about their child's behavior. Note that these and other questions were later published in a formal, validated tool intended to elicit information about parents' concerns – the PEDS [27].

The PEDS itself has demonstrated moderate internal consistency (Cronbach's alpha a = 0.692), with a tendency toward higher internal consistency as the age of participants increases [94]. The concurrent validity of the PEDS has been well studied, and associations between parent concerns on the PEDS and diagnostic measures of development are significant [94]. However, recent studies of the predictive validity of the tool are limited. Only reports of strong associations between parents' concerns and later observations of academic failure and diagnoses of autism spectrum disorder in children are reported by the tool's author [94].

3.3.3 Measuring DBPs: The Reference Standard

Following administration of the index test questions, the presence/absence of DBPs were confirmed using the Eyberg Child Behavior Inventory (ECBI) as a reference standard. The ECBI is a parent-report questionnaire comprised of 2 subscales: the problem subscale, and the intensity subscale. Relevant to this review, the problem scale asks parents to identify which behavior problems have been observed (if any) in their child, out of a total of 36 behaviors problems common among children with conduct problems. The ECBI was standardized on children between 2-12 years of age [95], which is consistent with this review's population of interest. The instrument has demonstrated good reliability, with an internal consistency coefficient (Cronbach's alpha) of r = 0.98 for the problem subscale [96]. Also, the ECBI has demonstrated good predictive validity, where children with conduct problems tend to have problem subscale scores (mean = 15.0) that are significantly higher than children without conduct problems (mean = 5.6, $p < 0.01$) [96]. Glascoe et al. defined the cut-off for a failed test (i.e., indicating the presence of DBPs) as the presence of 16 or more behavior problems [41]. A total of 20 children received a score of 16 or higher [41]. Unfortunately, due to the methods by which the results of the index test questions were pooled together, it is unknown which of the children with positive ECBI scores received a positive score on the first index question, and which of these children received a positive score for the second index test question.

3.4 Risk of Bias in Included Studies: Glascoe, MacLean & Stone, 1991

Table 3 summarizes the results of the risk of bias assessment completed for the study by Glascoe et al. [41] via the QUADAS-2 assessment tool. Two reviewers (SW & RB) independently completed the QUADAS-2 assessment for the included study. Appendix C contains a copy of the QUADAS-2 assessment form, including all signalling questions posed to reviewers for both risk of bias concerns and applicability concerns across all domains.

Table 3 Revised Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2): Results for the study by Glascoe, MacLean & Stone, 1991.

Risk of bias				Applicability concerns		
Patient selection	Index test	Reference standard	Flow and timing	Patient selection	Index test	Reference standard
☺	?	☹	☺	☺	☺	☺

☺ Low risk ☹ High Risk ? Unclear Risk

Note. Table adapted from the QUADAS-2 results template published by Whiting et al. [35].

3.4.1 Domain 1: Patient Selection

Risk of Bias in Patient Recruitment and Enrolment. The QUADAS-2 tool assessed whether the selection of patients by Glascoe et al. [41] could have introduced bias. Three signalling questions identified key methodological areas where risk of bias could have been introduced.

The first signalling question asked whether the study used "consecutive" or random sampling as a strategy to minimize the risk of sampling bias [36]. The study by Glascoe et al. used convenience sampling to enroll participants, where "parents of children seeking health services who were between 24 and 78 months of age and who were not acutely ill were asked to participate" [41]. As mentioned, all participants were parent-child dyads who had been recruited while seeking pediatric care from one of the five recruiting sites [41]. While convenience sampling carries some risk of sampling bias, reviewers did not find evidence in the text to suggest that participants were enrolled during the recruiting window in a manner that would impart additional risk of bias (e.g., cherry picking). Therefore, reviewers considered the risk of bias in this area to be low.

Second, the QUADAS-2 asked reviewers to determine whether a case-controlled design had been avoided by the study in question. The study by Glascoe et al. [41] used a naturalistic sampling design, where parents' concerns and children's behavior problems were assessed in tandem. The risk of bias pertaining to this question was therefore judged to be low.

Lastly, reviewers were asked to determine whether any unjustifiable exclusion criteria were implemented that could have introduced sampling bias. The study by Glascoe et al. had two explicit eligibility criteria. First, as mentioned, children were not eligible for inclusion if they were seeking acute medical care, and second, children were not eligible if they were younger than 24 months or older than 78 months of age [41]. These exclusion criteria were considered justifiable in the eyes of reviewers, given that the context of this review and of the practice guidelines in question [26] are also specific to well-child primary care visits; that is, routine, non-acute healthcare visits for young children (i.e., between 2-5 years of age).

One additional area not addressed by the QUADAS-2 tool may also contribute to the risk of bias in this study. 4 out of 99 recruited parent-child -dyads declined to participate in the study by Glascoe et al. [41]. Further, no data was reported by Glascoe et al. about the demographic or clinical features of these parent-child dyads to indicate that they differed significantly (or not) from those patients included in the study. Given the lack of transparency about potential differences between those individuals who participated and those who did not, reviewers determined that a risk of nonresponse bias [97] could not be ruled out. However, given that the return rate for this study was relatively high (i.e., responses gathered from 95/99 from eligible participants), reviewers considered

this risk to be low. Overall, risk of bias within the patient selection domain was judged to be low (see Table 3).

Applicability of the Patient Sample. The QUADAS-2 included a signalling question about whether the study included patients that did not match our review question. As mentioned, patients who participated in the study by Glascoe et al. [41] were sampled from pediatric care clinics and consisted of only those patients who were seeking non-acute care. Reviewers found that this clinical context was sufficiently similar to the clinical context detailed in the practice guidelines by Charach et al. [26], which informed our review questions. Therefore, reviewers had low concern about the overall applicability of the participants sampled by Glascoe et al. [41] (see Table 3).

3.4.2 Domain 2: Index Test

Risk of Bias within the Index Test. The QUADAS-2 tool also assessed whether the index test, how it was conducted, and how it was interpreted by Glascoe et al. [40] could have introduced bias. Two signalling questions assessed this. Reviewers were first asked to evaluate if the results of the index test used by Glascoe et al. [41] had been interpreted without knowledge of the results of the reference standard. In the study by Glascoe et al. [41], the index test questions were administered before the reference standard (the ECBI). Also, even if the results from the index test questions had been scored after the ECBI was administered, reviewers believe that the format of the index questions were consistent with those contained in the PEDS – that is, responses from parents are given by circling "yes" or "no", or "a little". In this format, parents' concerns can be assessed with relative objectivity, requiring little interpretation on behalf of researchers scoring the results. Therefore, reviewers considered the risk of bias pertaining to this question to be low.

Second, reviewers were asked to determine if a test threshold was used for the index test, and (if so) if the test's threshold had been pre-specified. As mentioned above, the index test used by Glascoe et al. [41] did not calculate or sum a total score for which a test threshold was necessary. Rather, as mentioned above, the total number of parents who reported concerns about their child's behavior was calculated by summing the number of participants who reported concerns about behavior in response to either the first (n = 8) or second (n = 26) index test question (Glascoe et al., 1991). Therefore, the risk of bias pertaining to this signalling question was found to be low.

Outside of the signalling questions asked by the QUADAS-2, however, reviewers identified additional areas of concern where the conduct of the index test may have introduced bias. If we assume that Glascoe et al. [41] used a version of the PEDS that is consistent with the published version of the tool as their index test there is some ambiguity in how Glascoe et al. [41] dichotomized parents' responses to the index test questions to arrive at that number of concerns (i.e., 34). The published version of the PEDS explicitly codes parents' responses to these same index test questions in one of three categories: "no", the parent is not concerned; the parent is "a little" concerned; or "yes" the parent is concerned [27]. If Glascoe et al. [41] used a similar scale to record index test responses, they may have combined responses of "yes" or "a little" to either index test question into one "yes" category for a dichotomous response variable. Any such strategy to determine what constituted a "yes" response on these index test questions may have influenced the total number of concerns measured, and thus poses a risk of bias. Therefore, the overall risk of bias in the index test domain was found to be unclear (see Table 3).

Applicability of the Index Test. The QUADAS-2 tool also included a signalling question that assessed whether the index test, its conduct, or its interpretation differed from our review questions. As mentioned previously, the second index test question posed by Glascoe et al. [41] specifically elicited parents' concerns about their child's behavior. As for the first index test question, only parents' responses that specifically referred to behavioral concerns were included in the analysis by Glascoe et al. [41]. Reviewers therefore had low concern about the applicability of the index test.

3.4.3 Domain 3: Reference Standard

Risk of Bias within the Reference Standard. The QUADAS-2 tool also asked reviewers to determine if the reference standard used by Glascoe et al. [41], its conduct, or its interpretation could have introduced bias. This was assessed via 2 signalling questions.

The first signalling question asked whether the reference standard was likely to classify the target condition correctly. According to the scoring guide that accompanies the QUADAS-2, "estimates of test accuracy are based on the assumption that the reference standard is 100% sensitive" [36]. To the knowledge of reviewers, there is no instrument that is 100% accurate in screening for behavior problems in young children. As such, any bias that results from the reference standard chosen by Glascoe et al. [41] should be assessed based on the reported validity and reliability of the tool itself.

Glascoe et al. [41] used the Eyberg Child Behavior Inventory (ECBI) as a reference standard. Thus, reviewers felt that the reference standard was likely to classify the target condition (behavior problems) correctly. Therefore, the risk of bias pertaining to the first signalling question was judged to be low.

The second signaling question asked reviewers to determine whether the reference standard results were interpreted without knowledge of the results of the index test. According to Glascoe et al., "after parents stated their concerns, they were asked to complete the Eyberg Child Behavior Inventory (ECBI)" [41]. The ECBI is designed such that the scoring of parent responses is intended to be objective, where scoring consists of summing the number of "yes" responses given by parents (for the problem scale) [98]. Therefore, there is little interpretation to be done on behalf of a researcher who scores the reference standard. Therefore, reviewers considered there to be a low risk of bias attributable to the interpretation of the reference standard results.

However, many additional aspects of the methods by which the ECBI was administered and scored remain unclear, and therefore represent potential sources of bias. First, it is unknown whether the same researcher administered both the index test and the ECBI to parents. Reviewers considered this to be the probable scenario, given that the time interval between tests was implied to be short. Also, it is unclear whether researchers assisted parents when they completed the ECBI. Hence, the effects of observer-expectancy bias on parents' responses to the reference standard cannot be ruled out.

Second, Glascoe et al. [41] defined a cutoff threshold for the ECBI of 16, where 16 or more behavior problems was considered a positive reference standard result. This cutoff, however, is inconsistent with the cutoff threshold defined by the authors of the ECBI, who have identified a "tentative cutoff point of 11" for the behavior problem scale [98]. Therefore, reviewers had concerns that the cutoff threshold for behavior problems that was used by Glascoe et al. [41] may

have biased (downward) the number of children identified as having behavior problems via the reference standard.

Lastly, the order of test administration could have introduced a response bias; that is, parents' previous responses to the index test may influence their later responses to the ECBI via a variety of mechanisms. The anchoring effect describes the tendency of an established reference point to influence an individual's decisions. For example, if a parent previously expressed that they did not have concerns about their child's behavior (i.e., via the index test), he or she may have been less likely to report the presence of specific behavior problems. Inversely, had the parents completed the ECBI first, a different type of response bias might have been introduced. Overall, reviewers considered there to be a high risk of bias within the reference standard domain (see Table 3).

Applicability of the Reference Standard. The QUADAS-2 tool asked reviewers to assess whether the target condition (i.e., DBPs), as defined by the reference standard used by Glascoe et al. [41] (i.e., the ECBI), matched the target condition defined in our review question. While the problems assessed by the ECBI may refer to behaviors common to many DBDs (i.e., ODD, Conduct Disorder, ADHD etc.), the authors of the ECBI only state that "the ECBI is an internally valid scale which measures a unitary construct, 'conduct problem'" [96]. Whether this tool possesses similar construct validity for our target condition (DBPs) is unknown. Reviewers considered the similarity between these two constructs to be sufficient such that there were low concerns about its applicability to the review questions.

3.4.4 Domain 4: Risk of Bias within Study Flow and Timing

Lastly, the QUADAS-2 asked reviewers to assess whether "patient flow could have introduced bias" [36]. The first signalling question asked reviewers if there was an appropriate time interval between the administration of the index test and of the reference standard. Reviewers were also asked to determine if any interventions were delivered between the administration of the index test and the reference standard. Glascoe et al. [41] simply state that "after parents stated their concerns [via the index test], they completed the ECBI", implying that both tests were administered within a short time interval and that no intervention occurred in between the two tests. Assuming this to be true, reviewers considered there to be a low risk of bias pertaining to this question.

In the second and third signaling questions, the QUADAS-2 asked reviewers to determine if all participants were assessed with a reference standard, and if all participants were assessed using the same reference standard. Only one reference standard was used (i.e., the ECBI) by Glascoe et al. [41]. Furthermore, data from all 95 individuals to whom the ECBI and the index test were administered were reported by Glascoe et al. [41]. Therefore, there were no missing data for either the reference standard or the index test. Reviewers therefore considered there to be a low risk of bias pertaining to these signaling questions.

The last signaling question asked reviewers to determine whether all data from patients enrolled in the study by Glascoe et al. [41] were included in their analysis. As mentioned, there was no evidence of missing data in the study by Glascoe et al. [41]. Therefore, reviewers considered there to be a low risk of bias pertaining to this question. Overall, reviewers found there to be a low risk of bias in the study flow and timing domain (see Table 3).

3.5 Results of Individual Studies: Glascoe, MacLean & Stone, 1991

Table 4 presents the frequencies of true positive, false negative, true negative, and false negative outcomes extracted from the study by Glascoe et al. [41]. Table 5 presents the proportions of true positive, false negative, true negative, and false positive outcomes, their standard errors, and 95% confidence intervals. There is an asymmetry in the number of parents who have concerns and the number of children who have DBPs, with more of the former than the latter. A notably larger proportion of parents expressed concerns about their child's behavior (0.358) than there were children who had behavioral problems (0.211). Indeed, the odds that a parent had concerns were 2.091 [i.e., (0.358/0.642)/(0.211/0.789)] times higher than the odds that a child had a DBP (Jackknife se = 0.058, Jackknife 95% CI: 1.977-2.205). Given this asymmetry, it is not wholly surprising that there is a relatively large proportion of false positive outcomes present in the sample (0.211). Also, there is a statistically significant proportion of false negative outcomes (0.063). Overall, parents made errors distinguishing between children with a DBP and children without in approximately 27.4% of cases.

Table 4 Frequencies of true positive, false positive, false negative, and true negative outcomes for the study by Glascoe, MacLean & Stone (1991).

		Behavioral Concerns		**Marginal Frequency**
		Yes	No	
Disruptive	Yes	14	6	20
Behavior Problems	No	20	55	75
Marginal Frequency		34	61	95

Table 5 Proportions of true positive, false positive, false negative, and true negative outcomes for the study by Glascoe, MacLean & Stone (1991).

		Behavioral Concerns		**Marginal Proportion**
		Yes	No	
Disruptive Behavior Problems	Yes	0.147 (0.036) [0.076-0.219]	0.063 (0.025) [0.014-0.112]	0.211 (0.042) [0.129–0.293]
	No	0.211 (0.042) [0.129-0.293]	0.579 (0.051) [0.480-0.678]	0.789 (0.042) [0.708–0.872]
Marginal Proportion		0.358 (0.049) [0.262-0.454]	0.642 (0.049) [0.546-0.739]	1.00

Note. Standard errors are given in parentheses; 95% confidence intervals are given in square parentheses.

3.6 Results of Data Synthesis & Analysis

Table 6 presents measures that indicate the accuracy with which parents' concerns can rule in the presence of a DBP in young children. An estimated PPV of 0.412 indicates that, among all parents who expressed concerns, 41.2% had a child with a DBP. An estimated SP of 0.733 indicates that, among all children without a DBP, 73.3% of their parents did not express behavioral concerns. RR1

was estimated to be 2.626, indicating that the odds that a child has a DBP are 2.6 times higher among parents who report concerns (i.e., 14/20 = 0.700), compared to the same odds irrespective of parents expressing concerns or not (i.e., 20/75 = 0.267). Following Kraemer [29] these measures were calibrated using a k(0,0) weighted kappa coefficient. An estimated k(0,0) value of 0.255 indicates only a fair agreement (i.e., $0.2 < k \leq 0.4$) between a positive reference standard result (i.e., presence of a DBP, as measured via the ECBI) and a positive index test result (i.e., presence of parents' concerns, as measured via the index test).

Table 6 Measures of the Accuracy of Parents' Concerns at Ruling In the Presence of a Disruptive Behavior Problem in Young Children: Results from Glascoe, MacLean & Stone (1991).

PPV	SP	RR1	k(0,0)
0.412 (0.084)	0.733 (0.051)	2.626 (0.068)	0.255 (0.009)
[0.246–0.577]	[0.633–0.833]	[2.493–2.760]	[0.238–0.272]

Note. PPV is the positive predictive value of parents' concerns; SP is the specificity of parents' concerns; Jackknife estimates of RR1 and k(0,0) are given, where RR1 is a relative risk ratio, and k(0,0) is a weighted kappa coefficient. Standard errors are given in parentheses; 95% confidence intervals are given in square parentheses.

Table 7 presents measures that indicate the accuracy with which parents' concerns can rule out the presence of a DBP in young children. An estimated NPV of 0.902 indicates that, among all parents who did not express concerns, 90.2% had a child without a DBP. An estimated SE of 0.7 indicates that, among all children with a DBP, 70% of their parents expressed behavioral concerns. RR3 was estimated to be 2.448, indicating that the odds that a child does *not* have a DBP are 2.5 times higher among parents that did *not* report concerns (i.e., 55/6 = 9.17), compared to the same odds irrespective of the presence or absence of parents' concerns (i.e., 75/20 = 3.75). Again, following Kraemer (1992), these measures were calibrated using a k(1,0) weighted kappa coefficient. An estimated k(1,0) value of 0.533 indicates only a moderate agreement (i.e., $0.4 < k \leq 0.6$) between a negative reference standard result (i.e., absence of a DBP, as measured via the ECBI) and a negative index test result (i.e., absence of parents' concerns, as measured via the index test).

Table 7 Measures of the Accuracy of Parents' Concerns at Ruling Out the Presence of a Disruptive Behavior Problem in Young Children: Results from Glascoe, MacLean & Stone (1991).

NPV	SE	RR3	k(1,0)
0.902 (0.038)	0.700 (0.102)	2.448 (0.100)	0.533 (0.01)
[0.827–0.976]	[0.499–0.901]	[2.252–2.644]	[0.501–0.564]

Note. NPV is the negative predictive value of parents' concerns; SE is the sensitivity of parents' concerns; Jackknife estimated of RR3 and k(1,0) are given, where RR3 is a relative risk ratio; k(1,0) is a weighted kappa coefficient. Standard errors are given in parentheses, and 95% confidence intervals are given in square parentheses.

4. Discussion

4.1 Overview

The purpose of this review was to determine whether parents' concerns can provide enough accurate information to PCPs, such that they can efficiently decide in favour of or against screening for DBPs in young children. We therefore sought to generate a consensus on the accuracy of parents' concerns in this context via systematic review methods. Ultimately, a limited number of results were retrieved by this review. The single study that met our eligibility criteria provided some preliminary evidence that parents' concerns *do not* accurately distinguish between children with DBPs and those without. Further, the results of our data analysis suggest that the presence of parents' concerns about their child's behavior may be more likely to generate false positive cases than false negatives – a trend which may negatively influence the accessibility of testing and diagnostic services for both children with DBPs and those without.

However, the results of one study are not sufficient to generate a consensus on this topic, and extreme caution should be exercised when interpreting and applying these results to clinical practice. The results of this review highlight a significant gap in the literature, which could indicate that high-quality evidence to demonstrate the accuracy of parents' concerns is simply difficult to retrieve using systematic review methods. Whether this is due to an overall dearth of research on this topic, or due to significant heterogeneity in the methods by which researchers have previously addressed this question remains unknown. The results of this review also prompt future research questions about whether alternate methods exist that can more accurately determine when screening for DBPs in young children is warranted.

These findings also call into question the legitimacy of current practice guidelines, which recommend: first, PCP's use the presence/absence of parents' concerns about their child's behavior to determine whether screening for DBPs is needed; and second, which do not make recommendations about the methods by which PCP's should elicit these concerns (Charach et al., 2017). The gap in literature illuminated by this review ultimately calls into question whether sufficient academic consensus is available to support these current recommendations.

4.2 Efficiency of Systematic Review Methodology: Scope and Applications

Our systematic search strategy produced 8420 records and resulted in 4964 abstracts screened for eligibility by reviewers. Just 53 of these records were eligible to be reviewed in full. Out of the 53 full-text studies reviewed, only one study [41] met all eligibility criteria. Despite the large number of records generated by our systematic search strategy, this review yielded more limited data than was previously identified in a recent rapid review [99]. Interestingly, most records that were excluded during full-text screening were excluded due to problems within the studies' designs (n = 33). Further, most of these records were excluded due to heterogeneity in how studies defined or measured parents' concerns and DBPs in children (*n* = 24/33). This result points to a need for additional research that a) generates consensus on how behavior problems and parents' concerns, as constructs, are defined, and b) develops gold standard screening tools that reliably measure these constructs. Until such time that behavioral concerns and DBPs are defined and measured consistently within the literature, systematic review methods may continue to be inefficient in retrieving data that addresses these constructs.

Given the results of this review, it is possible that diagnostic test accuracy methods may not be the most important method by which we can assess the predictive validity of parents' concerns. Instead, we may benefit from more observational data that assesses the longitudinal trajectory of FP and FN cases. Such work could provide much needed, objective data about the harms associated with both outcomes, and may provide an empirical basis on which we can identify which outcomes should be addressed with greater political attention. Alternatively, with unlimited resources, one could study the clinical outcomes associated with both FP and FN cases via RCT; however, such a method is likely not be feasible given ethical constraints.

4.3 The Accuracy of Parent's Concerns: Quality and Scope of Evidence

Ultimately, reviewers did not find good-quality evidence in the results of Glascoe et al. [41] to demonstrate that parents' concerns can accurately distinguish between children with DBPs and those without. While the reported sensitivity of parents' concerns (SE = 0.700, [0.499; 0.901]) was relatively high, calibration of this result demonstrated only moderate agreement (k(1,0) = 0.533, [0.501; 0.564]) between a negative reference standard result (i.e., absence of DBPs) and a negative index test result (i.e., absence of parents' concerns). Given that the estimated k(1,0) value is less than 0.6, this result suggests that the absence of parents' concerns may not provide enough accurate information to PCPs to rule *out* the presence of a DBP and decide against screening [36].

Similarly, the specificity of parents' concerns (SP = 0.733, [0.633; 0.833]) was also relatively high, but calibration of this result demonstrated only fair agreement (k(0,0) = 0.255, [0.238;0.272]) between a positive reference standard result (i.e., presence of DBPs) and a positive index test result (i.e., presence of parents' concerns). Given that the estimated value of k(0,0) is again less than 0.6, this result also suggests that the presence of parents' concerns may not provide enough accurate information to PCPs to rule *in* the presence of DBPs and decide in favour of screening.

Given the ratio of k(1,0) to k(0,0) (i.e., 0.533/0.255 = 2.091), parents' concerns appear to be relatively better at ruling *out* the presence of DBPs than they are at ruling *in*. In fact, any test that is better at ruling *out* illness will have a value of k(1,0) that is higher than k(0,0). When interpreting the relationship between these quality indices, it is important to assess the relative harms associated with a low value of either quality index. These harms are influenced by the relative prevalence of concerns (i.e., P and P'), and by the SE and SP of the test. As mentioned, while the SE of parents' concerns (0.700) was only slightly smaller than the SP (0.733), the quality of the specificity (k(0,0) = 0.255) was lower than that of the sensitivity (k(1,0) = 0.533). Additionally, there are more children without DBPs (P' = 0.789) in the sample from Glascoe et al. [40] than there are children with DBPs (P = 0.211), and more FP cases (i.e., 0.211 or 20/95) than FN cases (i.e., 0.063 or 6/95). Since the relative prevalence of P to P' are so disparate, and the SE and SP of parents' concerns are roughly equal, parents' concerns in this context are more likely to generate FP outcomes than FN outcomes.

There are significant harms associated with the use of a test that is relatively poor at ruling *in* illness (i.e., with a low k(0,0) value) and that which is likely to generate a high volume of FPs. This is due to the relative number of children affected by these kinds of errors, where more well children are likely to be misclassified as having a DBP. These harms can include unnecessary testing and increased emotional burden for a larger proportion of children without DBPs and their families. Indirectly, this scenario can also lead to harm for the smaller proportion of children with DBPs,

where a greater demand for screening and diagnostic services can reduce the accessibility of diagnostic testing for children who really need it. Thus, the harms associated with using a test that has a relatively poor ability to rule *in* illness (i.e., low k(0,0)) are likely to impact a majority of children, particularly in a system where there are finite resources available at a diagnostic level. In the context of the data from Glascoe et al. [41], it is therefore concerning that the quality of the specificity of parents' concerns (k(0,0)) is so low.

This is not to say that there are no harms associated with the use of a test that is poor at ruling *out* illness (i.e., where k(1,0) < k(0,0)). In the context of the data from Glascoe et al. [41], parents' concerns were still unlikely to provide enough accurate information to rule *out* the presence of DBPs given that the value of k(1,0) was smaller than 0.6 [37]. A poor ability to rule *out* illness in this context can still cause harm for the minority children who do have DBPs. Given the significant burden of suffering associated with the presence of DBPs, failure to decide in favour of screening (when necessary) may cost children access to timely interventions – interventions which, importantly, are effective in attenuating the risks of long-term harm. While these costs impact the minority of children in this sample (i.e., children with DBPs, P = 0.211), the scale of these costs is substantial for those affected, and therefore should not be undervalued.

In addition to the results of our data analysis, reviewers found there to be a high risk of bias within the methods of Glascoe et al. [41], which limits the strength of the evidence presented. There remains some ambiguity in how exactly parents' concerns were elicited in the study by Glascoe et al. [41], and by what index test. This points, again, to ambiguity in how parents' behavioral concerns are defined and measured in this field. Further, the order of test administration, and ambiguity in researcher blinding methods also contribute to a high risk of bias within the reported findings. Lastly, a discrepancy exists between the cut-off threshold defined by Glascoe et al. [41] and the cut-off threshold defined by the authors of the reference standard (i.e., the ECBI). This may have biased downward the estimates of accuracy reported here. Overall, the quality of the evidence retrieved by this review further emphasizes the need for more high-quality research to address whether parents' concerns can accurately distinguish between children with DBPs and children without. Until such a time that good quality evidence on this subject is accessible via methods like systematic reviews, clinicians may remain without the necessary tools to determine whether the presence or absence of parents' concerns justifies deciding in favour of or against screening for DBPs in young children.

4.4 Limitations

4.4.1 Limitations Due to Construct Heterogeneity

There is a known limitation introduced by the way in which the threshold for parents' concerns is defined. As mentioned previously, defining parents' concerns too narrowly, such that many concerns are excluded that do map to behavioral problems, can impact the validity of marginal proportions (i.e., the relative prevalence of behavior problems, P and P'). This, in turn can artificially reduce an SE estimate, and may also lead to improper estimates of k(1,0) and of the marginal odds ratios. Any change to the marginal odds estimates will impact the ratio of kappa coefficients, which then can impact the interpretation of whether a test is better at ruling *in* or ruling *out* illness. While studies that elicited behavioral, social, or emotional concerns from parents were eligible for inclusion in this review, the working definition of parents' concerns that was implemented by

Glascoe et al. [41] remained quite narrow relative to the definition of concerns implemented by this review. In the study by Glascoe et al. [41], parent concerns about domains of development outside of their child's behavior were excluded from their analysis. Therefore, it is possible that the estimates of SE and k(1,0) generated from Glascoe et al.'s data [41] and reported here, may be skewed downward.

4.4.2 The Absence of a Gold-Standard

In addition, the absence of a gold standard from which we may verify the presence/absence of DBPs limits our ability to estimate the accuracy of any index test. Not only does the absence of a gold standard contribute to the variability in the methods by which any study assesses the presence of DBPs in young children, the use of a non- "gold standard" test in diagnostic test accuracy may yield biased estimates of the prevalence of cases and non-cases [100]. Further, over- or underestimation of the prevalence of the target condition will result in over- or underestimation of the reliability of an index test [100]. The absence of a gold-standard tool therefore limits the reliability of the accuracy estimates presented by this review (as generated from data by Glascoe et al. [41]). Reviewers caution that the results presented here should not be interpreted as absolute indicators of the accuracy of parents' concerns, but rather as *estimates* that can be biased by classification errors in both the reference standard and the index test.

4.4.3 Limitations Imposed on Search Strategy

Several imposed search strategy constraints significantly impacted the number of studies retrieved for review. While reviewers employed an intentionally broad initial search approach, the final search strategy restricted the age range of children to 0-5 years. While a recently published rapid review [99] did assess the accuracy of parents' concerns for a broader target population (0-12-years) and it did not yield many more included studies (n = 2), it is possible that the age restriction imposed on our search strategy could have limited the number of studies ultimately retrieved. Additionally, the final search strategy excluded ADHD-related behavioral dimensions from key words. Given the high comorbidity between DBDs and ADHD [2, 101], relevant diagnostic test accuracy studies may have been overlooked. The omission of grey literature searches further restricted potential findings. These methodological constraints could ultimately have limited our ability to generate summary estimates of accuracy as planned in the meta-analysis protocol.

4.5 Interdisciplinary Contributions to Research and Future Directions

The current CPS screening guidelines for DBPs promote the psychological health and development of young children through routine and early screening [26]. This prevention strategy bridges the fields of public health, psychology, and medicine. In the context of these guidelines and this review, the assessment of the accuracy of parents' concerns further employed research and analysis methods from each of these three fields. This interdisciplinary approach attempted to bridge some of the research gaps that exist between the assessment of diagnostic test accuracy in medicine, and the assessment of screening test accuracy in psychology and public health.

While the results of this review did not provide robust answers to our research questions, they uncovered a gap in the current literature that may be addressed by future, interdisciplinary research.

As the prevention of psychological disorders becomes more central to public health policy, the development of accurate screening measures that can detect early symptoms may be in the interests of stakeholders. This may be particularly important for disorders that have an early age of onset and are associated with a considerable burden of suffering, like DBPs. With respect to the findings of this review, future research may benefit from addressing the accuracy of parents' concerns in a broader context. A focus on broader age ranges, and other developmental domains may facilitate the construction of a better-defined framework from which parents' concerns about their children's development can be categorized. Further, additional research that builds on the work of Glascoe et al. [41] is needed to develop a summary estimate of the accuracy of parents' concerns about their child's behavior.

As it stands, while parents' concerns seem to be quite poor at ruling *in* the presence of DBPs, eliciting parents' concerns to gather information about the developmental status of children may still be useful in the context of selective developmental screening strategies. In fact, despite limitations to the reliability of parents' concerns that are demonstrated by these findings, previous research has demonstrated that the expected benefit conferred by selective screening strategies that use parents' concerns is still greater than the expected benefit conferred by an alternative, universal screening strategy [102].

However, there are a wide variety of methods by which parents' concerns could be elicited that may improve their reliability within such selective screening strategies. In the context of the current Canadian practice guidelines, policy makers could consider implementing more explicit recommendations that better define *how* parents' concerns should be elicited by PCPs. For example, parents' concerns have been shown to be more reliable when elicited systematically by PCPs (i.e., via a questionnaire like the PEDS) [94]. The inclusion of specific methods or tools to the current recommendations that assist clinicians in eliciting parents' concerns may help minimize screening errors and minimize variability from clinician to clinician.

Considering the evidence presented in this review, current guidelines could also better identify which children are most in need of screening by incorporating an if/and algorithm into their practice recommendations. The current guidelines first instruct PCPs to elicit parents' concerns about their child's behavior. Given that parents' concerns are relatively poor at ruling *in* the presence of DBPs, policy makers could (for example) recommend that PCPs follow up with a second question that has a relatively better ability to rule *in* DBPs (I.e., has a higher k(0,0) value) than do parents' concerns. Following this, PCPs could be instructed to proceed with screening only if they receive positive responses from parents on both questions. Future studies would benefit from further examining whether there are other metrics that are more accurate in their ability to rule *in* the presence of DBPs than parents' concerns, and whether the implementation of such a decision-making procedure into practice guidelines would enhance the reliability with which we can identify children in need of early screening and intervention.

5. Conclusions

This systematic review reveals a critical gap in the evidence supporting current screening practices for DBPs in young children. The single eligible study from our comprehensive search [41] provides low-quality evidence that parents' concerns lack sufficient accuracy to reliably rule out DBPs in young children and are particularly poor at ruling in their presence. This finding has

significant implications for current clinical practice guidelines that rely on parents' concerns to guide screening decisions.

From a population health perspective, our findings raise serious questions about the effectiveness of current screening recommendations. The poor ruling-in capacity of parents' concerns is especially problematic, as it may lead to both unnecessary screening of children without DBPs and reduced accessibility of diagnostic services for those truly in need. These inefficiencies in resource allocation could have substantial impacts on early intervention efforts, which are known to be most effective during the preschool years.

The stark paucity of studies that have replicated Glascoe et al.'s work [41] points to a fundamental weakness in the evidence base supporting current practice guidelines [26]. Without robust evidence validating the accuracy of parents' concerns, the effectiveness of selective screening strategies remains questionable. Future research to address these issues should prioritize:

1. Developing standardized methods for eliciting and defining parents' concerns
2. Conducting large-scale validation studies of screening decision protocols
3. Evaluating alternative approaches to identifying children who most need DBP screening

Until such evidence is available, healthcare providers should exercise considerable caution when using parents' concerns as the sole basis for screening decisions. These findings strongly suggest the need to reassess and potentially revise current pediatric practice guidelines to ensure more effective identification of children with DBPs.

Appendix

Appendix A: Example of Search Strategy Adaptation for Medline(Ovid) Database and Date Searched.

Medline(Ovid) – Mar. 22, 2022

1	Child, Preschool/	972275
2	(child* or kid? or toddler? or girl? or boy? or preschool* or nurser* or pre-school* or prekindergarten* or kindergarten* or pediat* or paediat*).ti,ab.	1847261
3	1 or 2	2237264
4	Child Behavior Disorders/or "attention deficit and disruptive behavior disorders"/or conduct disorder/	26367
5	((dysfunc* or problem* or disrupt* or defian* or dysregulate*) adj behav*).ti,ab.	10927
6	((disrupt* or defian*) adj2 problem*).ti,ab.	573
7	exp Aggression/	41898
8	(aggression or aggressive*).ti,ab.	230200
9	"low concern for others".ti,ab.	9
10	("callous? unemotional traits" or "callous, unemotional traits").ti,ab.	681
11	(temper adj tantrum*).ti,ab.	250
12	(disobedien* or non-complian* or "non compliant" or "non compliance").ti,ab.	7772
13	problem behavior/	3290
14	4 or 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12 or 13	288961

15	exp Diagnosis/or mass screening/	9051820
16	diagnosis.fs.	2796540
17	(screen* or test or tests or testing).ti,ab.	3247904
18	((measur* or screen*) adj2 tool*).ti,ab.	45726
19	15 or 16 or 17 or 18	11855852
20	((parent* or mother* or father* or guardian* or caregiver*) adj2 (concern* or worr* or identifi* or complain* or opinion* or observ* or screen* or report* or insight* or attunement)).ti,ab.	47763
21	3 and 14 and 19 and 20	**1782**

**A total of 1782 records were retrieved from Medline.*

Appendix B: Data extraction form for Studies with a Naturalistic Sampling Design

ORGANIZATIONAL DATA					
REF ID		REVIEWER DATE		CHECKED BY	
Author/year					
Journal/source					
Country of origin					
Publication type					
STUDY CHARACTERISTICS					
Sample size					
Number of participants excluded					
Recruitment method					
Setting					

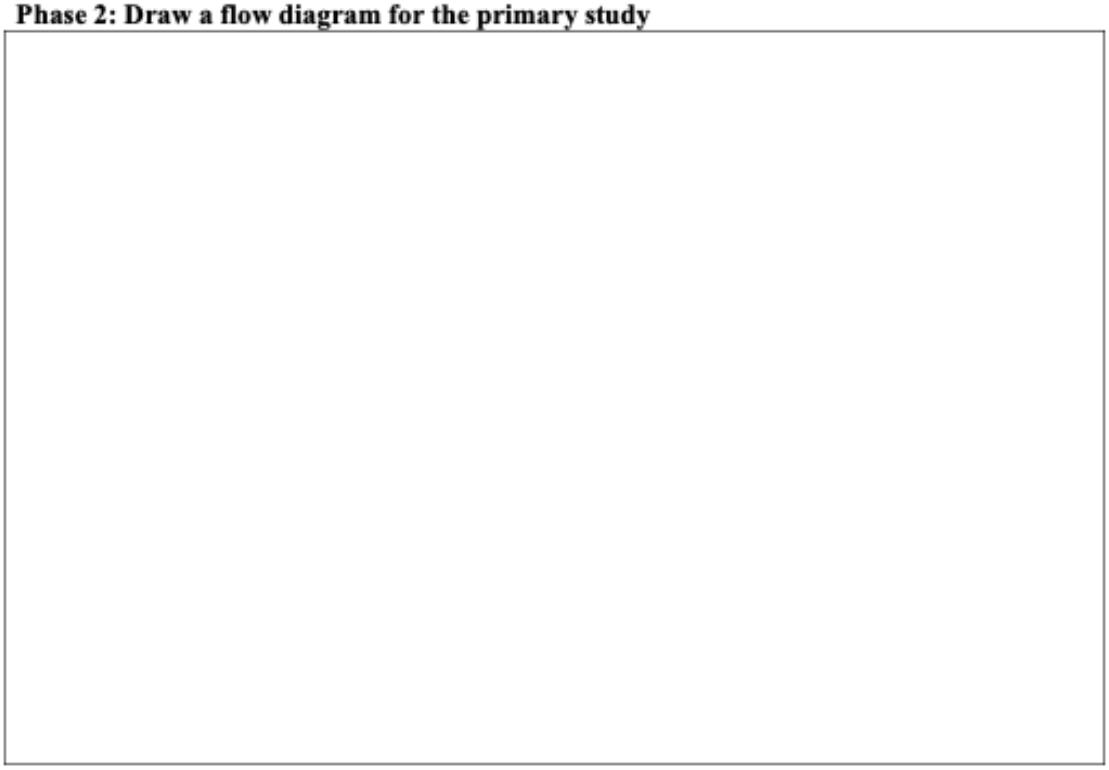
Location	
Funding	
Attrition	
PARTICIPANT CHARACTERISTICS (if reported)	
Age (mean, sd)	
Ethnicity (%)	
Sex (%)	
Annual household income	
Diagnostic status of participants at rdx (pre-study)	
Comorbid conditions?	
METHODOLOGY	
Behavioural problem(s) studied	
Index test used	
Reference standard used	
Timeline/flow of assessments	
Outcomes measures reported/assessed	

Appendix C: Quality Assessment of Diagnostic Accuracy Studies – 2 (QUADAS-2)

Phase 1: State the review question:

Patients (setting, intended use of index test, presentation, prior testing):
Index test(s):
Reference standard and target condition:

Phase 2: Draw a flow diagram for the primary study



Phase 3: Risk of bias and applicability judgments

QUADAS-2 is structured so that 4 key domains are each rated in terms of the risk of bias and the concern regarding applicability to the research question (as defined above). Each key domain has a set of signalling questions to help reach the judgments regarding bias and applicability.

DOMAIN 1: PATIENT SELECTION

A. Risk of Bias

Describe methods of patient selection:

- ❖ Was a consecutive or random sample of patients enrolled? Yes/No/Unclear
- ❖ Was a case-control design avoided? Yes/No/Unclear
- ❖ Did the study avoid inappropriate exclusions? Yes/No/Unclear

Could the selection of patients have introduced bias? RISK: LOW/HIGH/UNCLEAR

B. Concerns regarding applicability

Describe included patients (prior testing, presentation, intended use of index test and setting):

Is there concern that the included patients do not match the review question? CONCERN: LOW/HIGH/UNCLEAR

DOMAIN 2: INDEX TEST(S)

If more than one index test was used, please complete for each test.

A. Risk of Bias

Describe the index test and how it was conducted and interpreted:

- ❖ Were the index test results interpreted without knowledge of the results of the reference standard? Yes/No/Unclear
- ❖ If a threshold was used, was it pre-specified? Yes/No/Unclear

Could the conduct or interpretation of the index test have introduced bias? RISK: LOW /HIGH/UNCLEAR

B. Concerns regarding applicability

Is there concern that the index test, its conduct, or interpretation differ from the review question? CONCERN: LOW /HIGH/UNCLEAR

DOMAIN 3: REFERENCE STANDARD

A. Risk of Bias

Describe the reference standard and how it was conducted and interpreted:

- ❖ Is the reference standard likely to correctly classify the target condition? Yes/No/Unclear
- ❖ Were the reference standard results interpreted without knowledge of the results of the index test? Yes/No/Unclear

Could the reference standard, its conduct, or its interpretation have introduced bias? **RISK: LOW /HIGH/UNCLEAR**

B. Concerns regarding applicability

Is there concern that the target condition as defined by the reference standard does not match the review question? **CONCERN: LOW /HIGH/UNCLEAR**

DOMAIN 4: FLOW AND TIMING

A. Risk of Bias

Describe any patients who did not receive the index test(s) and/or reference standard or who were excluded from the 2x2 table (refer to flow diagram):

Describe the time interval and any interventions between index test(s) and reference standard:

- ❖ Was there an appropriate interval between index test(s) and reference standard? Yes/No/Unclear
- ❖ Did all patients receive a reference standard? Yes/No/Unclear
- ❖ Did patients receive the same reference standard? Yes/No/Unclear
- ❖ Were all patients included in the analysis? Yes/No/Unclear

Could the patient flow have introduced bias? **RISK: LOW /HIGH/UNCLEAR**

Acknowledgments

I would first like to express my gratitude to my thesis supervisor, Dr. Raymond Baillargeon, for his guidance throughout my time as his student. I am deeply appreciative of his encouragement and support through the completion of this project amidst some very uncertain and difficult times. I would also like to thank my thesis advisory committee members, Dr. Matthew McInnes and Dr. Philippe Robaey, for their expertise and advice throughout the duration of this project.

I would also like to express my thanks to the individuals who participated as members of the review team: Irfan Manji, Tanita Cepalo, Holly Paglia, Nihal Yapici, Raymond Baillargeon. Their contributions to each stage of the review process were instrumental in the completion of this work. Additionally, thank you to our research librarian, Nigèle Langlois, for her invaluable expertise in the development and adaptation of our search strategies.

Author Contributions

Sarah Wells wrote and completed all components of the research article and related research. Raymond Baillargeon co-wrote and edited all components of the research article.

Funding

The authors received no funding for the research or authorship of this review.

Competing Interests

The authors have declared that no competing interests exist.

References

1. Carter SA, Gray SA, Baillargeon RH, Wakschlag LS. A multidimensional approach to disruptive behaviors: Informing life span research from an early childhood perspective. In: Disruptive behavior disorders, advances in development and psychopathology: Brain research symposium series. New York: Springer Science + Business Media; 2013. pp. 103-135.
2. American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 5th ed. Arlington, VA: American Psychiatric Association; 2013. 461p.
3. Copeland WE, Shanahan L, Costello J, Angold A. Childhood and adolescent psychiatric disorders as predictors of young adult disorders. Arch Gen Psychiatry. 2009; 66; 764-772.
4. Caye A, Spadini AV, Karam RG, Grevet EH, Rovaris DL, Bau CH, et al. Predictors of persistence of ADHD into adulthood: A systematic review of the literature and meta-analysis. Eur Child Adolesc Psychiatry. 2016; 25: 1151-1159.
5. Charach A, Mohammadzadeh F, Bélanger SA, Easson A, Lipman EL, McLennan JD, et al. Identification of preschool children with mental health problems in primary care: Systematic review and meta-analysis. J Can Acad Child Adolesc Psychiatry. 2020; 29: 76-105.
6. Waddell C, McEwan K, Shepherd C, Offord D, Hua J. A public health strategy to improve the mental health of Canadian children. Can J Psychiatry. 2005; 50: 226-233.
7. Waddell C, McEwan K, Peters RD, Hua JM, Garland O. Preventing mental disorders in children: A public health priority. Can J Public Health. 2007; 98: 174-178.
8. Offord DR, Kraemer HC, Kazdin AE, Jensen PS, Harrington R, Gardner JS. Lowering the burden of suffering: Monitoring the benefits of clinical, targeted, and universal approaches. In: Developmental health and the wealth of nations. New York, NY: The Guilford Press; 1999. pp. 293-310.
9. Barican JL, Yung D, Schwartz C, Zheng Y, Georgiades K, Waddell C. Prevalence of childhood mental disorders in high-income countries: A systematic review and meta-analysis to inform policymaking. BMJ Ment Health. 2022; 25: 36-44.

10. Polanczyk GV, Salum GA, Sugaya LS, Caye A, Rohde LA. Annual research review: A meta-analysis of the worldwide prevalence of mental disorders in children and adolescents. J Child Psychol Psychiatry. 2015; 56: 345-365.
11. Offord DR, Bennett KJ. Conduct disorder: Long-term outcomes and intervention effectiveness. J Am Acad Child Adolesc Psychiatry. 1994; 33: 1069-1078.
12. Erskine HE, Norman RE, Ferrari AJ, Chan GC, Copeland WE, Whiteford HA, et al. Long-term outcomes of attention-deficit/hyperactivity disorder and conduct disorder: A systematic review and meta-analysis. J Am Acad Child Adolesc Psychiatry. 2016; 55: 841-850.
13. Scott JG, Pedersen MG, Erskine HE, Bikic A, Demontis D, McGrath JJ, et al. Mortality in individuals with disruptive behavior disorders diagnosed by specialist services–A nationwide cohort study. Psychiatry Res. 2017; 251: 255-260.
14. Reardon T, Harvey K, Baranowska M, O'brien D, Smith L, Creswell C. What do parents perceive are the barriers and facilitators to accessing psychological treatment for mental health problems in children and adolescents? A systematic review of qualitative and quantitative studies. Eur Child Adolesc Psychiatry. 2017; 26: 623-647.
15. Christenson JD, Crane DR, Malloy J, Parker S. The cost of oppositional defiant disorder and disruptive behavior: A review of the literature. J Child Fam Stud. 2016; 25: 2649-2658.
16. Foster EM, Jones DE, Conduct Problems Prevention Research Group. The high costs of aggression: Public expenditures resulting from conduct disorder. Am J Public Health. 2005; 95: 1767-1772.
17. Matza LS, Paramore C, Prasad M. A review of the economic burden of ADHD. Cost Eff Resour Alloc. 2005; 3: 5.
18. Leibson CL, Long KH. Economic implications of attention-deficit hyperactivity disorder for healthcare systems. Pharmacoeconomics. 2003; 21: 1239-1262.
19. Jekel JF, Katz DL, Elmore JG, Wild DM. Understanding the quality of data in clinical medicine. In: Epidemiology, biostatistics, and preventive medicine. 3rd ed. Philadelphia, PA: Saunders Elsevier; 2007. pp. 108-114.
20. Chacko A, Granski M, Horn EP, Levy MD, Dahl V, Lacks RS, et al. Prevention of disruptive behavior problems in children. In: Developmental pathways to disruptive, impulse-control and conduct disorders. New York, NY: Academic Press; 2018. pp. 347-380.
21. Bayer J, Hiscock H, Scalzo K, Mathers M, McDonald M, Morris A, et al. Systematic review of preventive interventions for children's mental health: What would work in Australian contexts? Aust N Z J Psychiatry. 2009; 43: 695-710.
22. Lipkin PH, Macias MM, Norwood KW, Brei TJ, Davidson LF, Davis BE, et al. Promoting optimal development: Identifying infants and young children with developmental disorders through developmental surveillance and screening. Pediatrics. 2020; 145: e20193449.
23. Workgroup BF, Richerson JE, Simon GR, Abularrage JJ, Boudreau AD, Baker CN, et al. 2017 recommendations for preventive pediatric health care. Pediatrics. 2017; 139: e20170254.
24. Baillargeon RH, Keenan K, Cao G. The development of opposition-defiance during toddlerhood: A population-based cohort study. J Dev Behav Pediatr. 2012; 33: 608-617.
25. Petitclerc A, Tremblay RE. Childhood disruptive behaviour disorders: Review of their origin, development, and prevention. Can J Psychiatry. 2009; 54: 222-231.

26. Charach A, Bélanger SA, McLennan JD, Nixon MK. Le dépistage des comportements perturbateurs en première ligne chez les enfants d'âge préscolaire. Paediatr Child Health. 2017; 22: 478-493.
27. Glascoe F. Parents' evaluations of developmental status: A method for detecting and addressing developmental and behavioral problems in children. Nolensville, TN: Ellsworth & Vandermeer Press; 1997.
28. Eyberg SM. Eyberg child behavior inventory and Sutter-Eyberg student behavior inventory-revised: Professional manual. Odessa, FL: Phychological Assessment Resources; 1999.
29. Kraemer HC. Evaluating medical tests: Objective and quantitative guidelines. Newbury Park, CA: SAGE Publications; 1992.
30. Kraemer HC, Gibbons RD. Where do we go wrong in assessing risk factors, diagnostic and prognostic tests? The problems of two-by-two association. Psychiatr Ann. 2009; 39: 711-718.
31. Covidence systematic review software. Veritas Health Innovation [Internet]. Melbourne, Australia: Covidence systematic review software. Available from: www.covidence.org.
32. Briscoe S, Bethel A, Rogers M. Conduct and reporting of citation searching in Cochrane systematic reviews: A cross-sectional study. Res Synth Methods. 2020; 11: 169-180.
33. The Cochrane Collaboration. Data Collection form for intervention review – RCTs and non-RCTs [Internet]. London, UK: The Cochrane Collaboration; 2020 [cited date 2020 December 4]. Available from: https://dplp.cochrane.org/data-extraction-forms.
34. McInnes MD, Moher D, Thombs BD, McGrath TA, Bossuyt PM, Clifford T, et al. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: The PRISMA-DTA statement. JAMA. 2018; 319: 388-396.
35. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011; 155: 529-536.
36. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2 background document. Bristol: University of Bristol; 2011.
37. Landis JR. The measurement of observer agreement for categorical data. Biometrics. 1977; 33: 159-174.
38. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol. 2005; 58: 982-990.
39. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ. 2021; 372: n71.
40. Melo I, Rodriguez R. Correlations between present disorders noted in the records of school-age children and disorders of early childhood mentioned by parents during medical history taking. Rev Med Suisse Romande. 1980; 100: 267-271.
41. Glascoe FP, MacLean WE, Stone WL. The importance of parents' concerns about their child's behavior. Clin Pediatr. 1991; 30: 8-11.
42. Reijneveld SA, de Meer G, Wiefferink CH, Crone MR. Parents' concerns about children are highly prevalent but often not confirmed by child doctors and nurses. BMC Public Health. 2008; 8: 124.
43. August GJ, Realmuto GM, Crosby RD, MacDonald AW. Community-based multiple-gate screening of children at risk for conduct disorder. J Abnorm Child Psychol. 1995; 23: 521-544.

44. Berger-Jenkins E, Monk C, D'Onfro K, Sultana M, Brandt L, Ankam J, et al. Screening for both child behavior and social determinants of health in pediatric primary care. J Dev Behav Pediatr. 2019; 40: 415-424.
45. Bengi-Arslan L, Verhulst FC, van der Ende J, Erol N. Understanding childhood (problem) behaviors from a cultural perspective: Comparison of problem behaviors and competencies in Turkish immigrant, Turkish and Dutch children. Soc Psychiatry Psychiatr Epidemiol. 1997; 32: 477-484.
46. Bergold S, Christiansen H, Steinmayr R. Interrater agreement and discrepancy when assessing problem behaviors, social-emotional skills, and developmental status of kindergarten children. J Clin Psychol. 2019; 75: 2210-2232.
47. Briggs-Gowan MJ. A parent assessment of social-emotional and behavior problems and competence for infants and toddlers: Reliability, validity, and associations with maternal symptoms and parenting stress. New Haven, CT: Yale University; 1996.
48. Briggs-Gowan MJ, Carter AS, McCarthy K, Augustyn M, Caronna E, Clark R. Clinical validity of a brief measure of early childhood social–emotional/behavioral problems. J Pediatr Psychol. 2013; 38: 577-587.
49. Campbell SB, Breaux AM, Ewing LJ, Szumowski EK. A one-year follow-up study of parent-refer red hyperactive preschool children. J Am Acad Child Psychiatry. 1984; 23: 243-249.
50. Holtz CA. Screening of behavior problems in young children from low-income families: The development of a new assessment tool. Milwaukee, WI: Marquette University; 2010.
51. Lavigne JV, Bryant FB, Hopkins J, Gouze KR. Age 4 predictors of oppositional defiant disorder in early grammar school. J Clin Child Adolesc Psychol. 2019; 48: 93-107.
52. Nærde A, Ogden T, Janson H, Zachrisson HD. Normative development of physical aggression from 8 to 26 months. Dev Psychol. 2014; 50: 1710-1720.
53. Plamondon A, Browne DT, Madigan S, Jenkins JM. Disentangling child-specific and family-wide processes underlying negative mother-child transactions. J Abnorm Child Psychol. 2018; 46: 437-447.
54. Sabol TJ, Kessler CL, Rogers LO, Petitclerc A, Silver J, Briggs-Gowan M, et al. A window into racial and socioeconomic status disparities in preschool disciplinary action using developmental methodology. Ann N Y Acad Sci. 2022; 1508: 123-136.
55. Wakschlag LS, Choi SW, Carter AS, Hullsiek H, Burns J, McCarthy K, et al. Defining the developmental parameters of temper loss in early childhood: Implications for developmental psychopathology. J Child Psychol Psychiatry. 2012; 53: 1099-1108.
56. Wakschlag LS, Briggs-Gowan MJ, Choi SW, Nichols SR, Kestler J, Burns JL, et al. Advancing a multidimensional, developmental spectrum approach to preschool disruptive behavior. J Am Acad Child Adolesc Psychiatry. 2014; 53: 82-96.
57. Winsper C, Wolke D. Infant and toddler crying, sleeping and feeding problems and trajectories of dysregulated behavior across childhood. J Abnorm Child Psychol. 2014; 42: 831-843.
58. August GJ, Braswell L, Thuras P. Diagnostic stability of ADHD in a community sample of school-aged children screened for disruptive behavior. J Abnorm Child Psychol. 1998; 26: 345-356.
59. Poduska JM. Parents' perceptions of their first graders' need for mental health and educational services. J Am Acad Child Adolesc Psychiatry. 2000; 39: 584-591.
60. Feldman MA, Hancock CL, Rielly N, Minnes P, Cairns C. Behavior problems in young children with or at risk for developmental delay. J Child Fam Stud. 2000; 9: 247-261.

61. Fanton JH, MacDonald B, Harvey EA. Preschool parent-pediatrician consultations and predictive referral patterns for problematic behaviors. J Dev Behav Pediatr. 2008; 29: 475-482.
62. Glascoe FP, Bell RT. Behavioural and developmental problems in children. Patient Care. 1998; 32: 63-72.
63. Gray SA, Carter AS, Briggs-Gowan MJ, Hill C, Danis B, Keenan K, et al. Preschool children's observed disruptive behavior: Variations across sex, interactional context, and disruptive psychopathology. J Clin Child Adolesc Psychol. 2012; 41: 499-507.
64. Herman-Staab B. Screening, management, and appropriate referral for pediatric behavior problems. Nurse Pract. 1994; 19: 40-42.
65. Ireton H. The child development review: Monitoring children's development using parents' and pediatricians' observations. Infants Young Child. 1996; 9: 42-52.
66. Ogden T, Hagen KA. Treatment effectiveness of parent management training in Norway: A randomized controlled trial of children with conduct problems. J Consult Clin Psychol. 2008; 76: 607-621.
67. Szczepaniak D, McHenry MS, Nutakki K, Bauer NS, Downs SM. The prevalence of at-risk development in children 30 to 60 months old presenting with disruptive behaviors. Clin Pediatr. 2013; 52: 942-949.
68. Glascoe FP. A method for deciding how to respond to parents' concerns about development and behavior. Ambul Child Health. 1999; 5: 197-208.
69. Glascoe FP. Parents' evaluation of developmental status: How well do parents' concerns identify children with behavioral and emotional problems? Clin Pediatr. 2003; 42: 133-138.
70. Barkley RA, Shelton TL, Crosswait C, Moorehouse M, Fletcher K, Barrett S, et al. Preschool children with disruptive behavior: Three-year outcome as a function of adaptive disability. Dev Psychopathol. 2002; 14: 45-67.
71. Karabekiroglu K, Uslu R, Kapci-Seyitoglu EG, Özbaran B, Öztop DB, Özel-Özcan Ö, et al. A nationwide study of social–emotional problems in young children in Turkey. Infant Behav Dev. 2013; 36: 162-170.
72. Kruizinga I, Jansen W, de Haan CL, Raat H. Reliability and validity of the KIPPPI: An early detection tool for psychosocial problems in toddlers. PloS One. 2012; 7: e49633.
73. Wendland J, Danet M, Gacoin E, Didane N, Bodeau N, Saïas T, et al. French version of the brief infant-toddler social and emotional assessment questionnaire–BITSEA. J Pediatr Psychol. 2014; 39: 562-575.
74. Alakortes J, Kovaniemi S, Carter AS, Bloigu R, Moilanen IK, Ebeling HE. Do child healthcare professionals and parents recognize social-emotional and behavioral problems in 1-year-old infants? Eur Child Adolesc Psychiatry. 2017; 26: 481-495.
75. Briggs-Gowan MJ, Carter AS. Social-emotional screening status in early childhood predicts elementary school outcomes. Pediatrics. 2008; 121: 957-962.
76. Chen IC, Lee HC, Yeh GC, Lai CH, Chen SC. The relationship between parental concerns and professional assessment in developmental delay in infants and children--a hospital-based study. J Chin Med Assoc. 2004; 67: 239-244.
77. Ellingson KD, Briggs-Gowan MJ, Carter AS, Horwitz SM. Parent identification of early emerging child behavior problems: Predictors of sharing parental concern with health providers. Arch Pediatr Adolesc Med. 2004; 158: 766-772.

78. Godoy L, Carter AS, Silver RB, Dickstein S, Seifer R. Infants and toddlers left behind: Mental health screening and consultation in primary care. J Dev Behav Pediatr. 2014; 35: 334-343.
79. Sheldrick RC, Neger EN, Perrin EC. Concerns about development, behavior, and learning among parents seeking pediatric care. J Dev Behav Pediatr. 2012; 33: 156-160.
80. Barkauskienė R, Bongarzoni AD, Bieliauskaitė R, Jusienė R, Raižienė S. Attention-deficit/hyperactivity disorder: Possibilities of early diagnostics. Medicina. 2009; 45: 764-771.
81. Feeney-Kettler KA. Early identification of preschool students at risk for emotional and behavioral disorders: Development and validation of a parent-teacher screener. Madison, WI: University of Wisconsin--Madison; 2008.
82. Lorber MF, Del Vecchio T, Slep AM. Infant externalizing behavior as a self-organizing construct. Dev Psychol. 2014; 50: 1854-1861.
83. Sand EA. Symptômes de comportement d'enfants belges observés par leur mère II: Associations entre symptômes (3 à 9 ans). Rev Neuropsychiatr Infantile D'Hygiène Ment L'Enfance. 1972; 20: 253-265.
84. Sawyer MG, Mudge J, Carty V, Baghurst P, McMichael A. A prospective study of childhood emotional and behavioural problems in Port Pirie, South Australia. Aust N Z J Psychiatry. 1996; 30: 781-787.
85. Thomas BH, Byrne C, Offord DR, Boyle MH. Prevalence of behavioral symptoms and the relationship of child, parent, and family variables in 4− and 5-year-olds: Results from the Ontario child health study. J Dev Behav Pediatr. 1991; 12: 177-184.
86. Ware HS, Jouriles EN, Spiller LC, McDonald R, Swank PR, Norwood WD. Conduct problems among children at battered women's shelters: Prevalence and stability of maternal reports. J Fam Violence. 2001; 16: 291-307.
87. Alink LR, Mesman J, Van Zeijl J, Stolk MN, Juffer F, Koot HM, et al. The early childhood aggression curve: Development of physical aggression in 10-to 50-month-old children. Child Dev. 2006; 77: 954-966.
88. Edelstein ML, Moen A, Benson JL, Smucker R, Perkins-Parks S. Development and implementation of a function-based clinical interview to evaluate childhood behavior problems. Cogn Behav Pract. 2023; 30: 421-435.
89. Glascoe FP, Altemeier WA, MacLean WE. The importance of parents' concerns about their child's development. Am J Dis Child. 1989; 143: 955-958.
90. Glascoe FP. Using parents' concerns to detect and address developmental and behavioral problems. J Spec Pediatr Nurs. 1999; 4: 24-35.
91. Ilić SB, Nikolić SJ, Ilić-Stošović DD, Golubović ŠS. Early identification of children with developmental delay and behavioural problems according to parents concerns in the Republic of Serbia. Early Child Dev Care. 2020; 190: 2605-2611.
92. Pavuluri MN, Luk SL, Clarkson J, McGee R. A community study of preschool behaviour disorder in New Zealand. Aust N Z J Psychiatry. 1995; 29: 454-462.
93. Studts CR, Polaha J, van Zyl MA. Identifying unbiased items for screening preschoolers for disruptive behavior problems. J Pediatr Psychol. 2017; 42: 476-486.
94. Glascoe FP. The Validity of PEDS. In: Collaborating with parents: Using parents' evaluation of developmental status to detect and address developmental and behavioral problems. 2nd ed. PEDStest.com, LLC.; 2013. pp. 115-135.

95. Robinson EA, Eyberg SM, Ross AW. The standardization of an inventory of child conduct problem behaviors. J Clin Child Adolesc Psychol. 1980; 9: 22-28.
96. Glascoe FP. The Reliability of PEDS. In: Collaborating with parents: Using parents' evaluation of developmental status to detect and address developmental and behavioral problems. 2nd ed. PEDStest.com, LLC.; 2013. pp. 103-114.
97. Goodwin KA, Goodwin CJ. Research in psychology: Methods and design. In: Research in psychology: Methods and design. 5th ed. John Wiley & Sons; 2008. pp. 423-460.
98. Eyberg SM, Ross AW. Assessment of child behavior problems: The validation of a new inventory. J Clin Child Adolesc Psychol. 1978; 7: 113-116.
99. Baillargeon RH, Charette M, Tessier F, Brand KP. Clinical practice guidelines about screening for disruptive behavior problems at well-child visits: A rapid review of the literature on the accuracy of parents' behavioral concerns. OBM Integr Complement Med. 2022; 7: 031.
100. Baillargeon RH, Tremblay RE, Willms D, Lee KH, Romano E, Wu HX. Modeling intraindividual change over time in the absence of a "Gold Standard". Psychol Sci. 2004; 46: 7-34.
101. Loeber R, Burke JD, Pardini DA. Development and etiology of disruptive and delinquent behavior. Annu Rev Clin Psychol. 2009; 5: 291-310.
102. Baillargeon RH. Clinical practice guidelines for monitoring children's behavioural development at the 18-month well-baby visit: A decision analysis comparing the expected benefit of two alternative strategies. J Eval Clin Pract. 2021; 27: 62-68.