April 2024

Screening and Triage for Autism Assessment Pathways

A report of tools and recommendations for improved practice and service delivery

Executive summary

In England, waiting times for autism assessments for children, young people and adults commissioned or provided by the NHS are currently at record high levels. Within referral pathways, tools and questionnaires are used to inform clinical decisions around autism assessment, such as when to refer a person to a specialist autism service for assessment, what type of service to refer to, who to prioritise for assessment, and what type of assessment to conduct. In this report, decisions around whether to refer a person for an assessment or whether the assessment service should accept a referral are defined as **screening**. Decisions around which patients to prioritise for an assessment and identifying which components will be included in an autism assessment are defined as **triage**.

A range of tools have been developed to assist with screening and triage in the autism assessment pathway, although there is currently no resource available to support with appraising these. In the absence of this information, services risk using tools that are not validated or that are not appropriate for use in a particular context.

The key aim of this report was to determine which criteria should be used to evaluate standardised tools used for screening and triage in autism assessment pathways in England for children, young people and adults, and to evaluate example tools against these criteria. This report **defines the criteria**, outlines **considerations for applying these** in different use cases, and presents some **self-assessment questions** to guide decision making around which tools best fit the criteria.

By setting out criteria that can be applied consistently across services, we can strive to achieve a more uniform and systematic approach to tackling the diagnostic assessment waitlist crisis and providing standardised high quality assessment services. Services can utilise the criteria to decide which tools are most appropriate for their needs and to evaluate the tools that are currently being used. In this way we can ensure that all autistic people can be better supported to get the right support at the right time, and achieve happy, healthy, long lives.

Methods and findings

To identify the criteria, we convened stakeholders to define considerations about the use of screening instruments, triage processes, factors related to gender and intersectionality, and to establish best practice guidance about appraising the effectiveness of tools across use cases relevant to NHS autism assessment pathways. This included representation from an Expert Reference Group (ERG) of clinicians and researchers and lived experience views of autistic people and family members of autistic people.

The ERG selected example tools to be evaluated against the criteria defined in this report. A narrative literature review was also used to appraise the available evidence for these tools against the criteria. Due to the extensive literature and limited nature of discussions, not all tools which are in use or have been developed could be included in the report.

The criteria defined in this report are as follows:

- **Validity and reliability –** Validity refers to whether the tool measures what it intends to measure. Reliability refers to whether the output is consistent when repeated.
- **Acceptability –** Acceptability concerns how well a tool is received by patients, their families/carers, assessment teams and referrers.
- **Feasibility and efficiency –** Feasibility refers to whether it is practical and viable to use for a specific purpose. Efficiency refers to whether a tool is effective at its intended aims.
- **Extent of evidence –** This refers to how much evidence is available, both quality and quantity. Is important to consider the whole evidence base and whether findings have been consistently replicated by multiple studies.
- **Design characteristics of studies –** It is important to appraise the characteristics of studies to determine the risk of unintentional or intentional bias in how tools have been evaluated.
- **Outcomes –** Tools may be associated with different outcomes, including conversion rate, and the outputs they produce.
- **Flagging of alternative/co-occurring conditions –** Many neurodevelopmental conditions frequently co-occur with autism. There may be benefits to tools which can screen for alternative or co-occurring diagnoses.
- **Consistency across populations -** Tools should demonstrate good performance across populations and subgroups (including gender, age, ethnicity, and disability).

No tools scored highly on all specified criteria and most tools demonstrated mixed results across different criteria. There was limited evidence for the validity, reliability and feasibility of tools in the context of the NHS autism assessment pathway, and research design has not tended to reflect how they are used in practice by patients and services.

Standardised tools may continue to have utility in assisting with screening and triage, but the concerns around their validity and bias suggest that decision-making at these stages of the assessment pathway should be accompanied by clinical judgement of knowledgeable clinicians, rather than relying on scores on tools alone to decide to refer or prioritise people for assessment.

This report is made up of seven sections, as follows:

- **Section 1** provides background information on the issue,
- **Section 2** sets out the aims and scope of the project,
- **Section 3** describes the methods used in this project,
- **Section 4** outlines the findings,
- **Section 5** outlines different use cases in which tools may be used,
- **Section 6** presents the criteria recommended for appraising tools for their use in screening and triage, including self-assessment questions for providers to aid decision-making,
- **Section 7** presents a summary of the research literature on some of the tools used by services and suggests directions for future research.

April 2024

Contents

autistica

Screening and Triage for Autism Assessment Pathways

A review of tools and recommendations for improved practice and service delivery

1. Background

Around 1-2% of the population are estimated to be autistic; if everyone who is autistic were to be assessed we would expect that around 1.16% of children (Baird et al., 2006) and 1.1% of adults (Brugha et al., 2012) should receive an autism diagnosis. Autism is diagnosed at substantially different rates in different age cohorts in England, with apparent over-diagnosis in some younger cohorts and under-diagnosis in older people. The numbers of people that receive a diagnosis each year has been rapidly increasing over the past decade, likely due to historical under provision, under identification, changes in public and professional awareness, shifting diagnostic criteria, and assessment practices (O'Nions et al., 2023).

Waiting times for autism assessments for children, young people and adults commissioned or provided by the NHS in England are currently at record high levels. As of December 2023 (the latest data available at time of writing in March 2024), there were 172,022 people with an open referral, with 85.5% of these referrals open for at least 13 weeks, and new referrals increased nearly 50% on the previous year (NHS Digital, 2024).

In England, NHS services, including specialist autism assessment services and community mental health teams, carry out autism assessments for children, young people and adults, giving a diagnosis where appropriate. A person may be flagged as requiring an assessment when they self-present to services, have a family member enquire on their behalf, or are referred by a professional. Referrals to these services may originate from primary care providers (such as GPs), primary or secondary mental health services, education, and other areas such as the criminal justice system.

Within these referral pathways, tools and questionnaires are used to inform clinical decisions around autism assessment, such as:

- **For a referrer** – when to refer someone to a specialist autism team for an autism assessment, and when not to,
- **For a referrer** – whether to refer someone to a service that only conducts autism assessments, or if a referral to a service that can assess for and diagnose other possible conditions is warranted,
- **For an assessment provider** – to decide which of the people referred to the service should be offered an assessment, in what order people are assessed, and who should not be assessed,
- **For an assessment provider** – to decide what type of assessment(s) to undertake, and which clinical professional(s) should conduct the assessment.

In this report, decisions around whether to refer a person for an assessment or whether the assessment service should accept a referral are defined as **screening**. Decisions around prioritising which people should receive an autism assessment first and identifying which components will be included in an autism assessment are defined as **triage**. Outcomes of

screening and triage can include recommending an assessment, discharge from a service, an open-ended referral to allow for future review if needed, assessment for another intervention or service, or referral to a national specialist service (for complex presentations); interim support may be offered by other services while a person awaits assessment (NHS England, 2023b).

Autism assessment services face limited resources to conduct assessments and high demand for assessments. Services which refer to and carry out autism assessments are operated by different NHS Trusts, providers commissioned by Integrated Care Boards and/or Local Authorities depending on their locality. At present, these authorities can decide independently which tools to use for the scenarios listed above. A range of tools has been developed to assist with screening decisions and triage of referrals, including those which are standardised and validated. For example, the Autism Spectrum Quotient (AQ; and variations of the AQ) has undergone extensive testing (Baghdadli et al., 2017), leading to a strong evidence base for the strengths and weaknesses of this questionnaire's use as a tool as part of the process of screening for autism. Guidance from the National Institute for Health and Care Excellence (NICE) around autism diagnosis in adults (NICE, 2021) recommends the 10-item version of the AQ for this purpose, with a recommended cut-off score of 6 or above indicating that an assessment may be warranted (although note that a previous error with NICE guidance reported this at above 6; Waldren et al., 2021). NICE guidance also recommends the use of the Ritvo Asperger and Autism Diagnostic Scale-Revised (RAADS-R) as part of the assessment, but not in screening procedures. No specific tools are recommended by NICE (2017) in the diagnostic assessment process for children. NHS England (2023a) commissioning guidance states that processes should be standardised across services that provide diagnostic assessment.

Without easy access to information appraising tools used for screening and triage, services risk using instruments that are not validated or that are not appropriate for use in a particular context. For example, in the absence of appropriate guidance, an instrument that is designed to give an overview of a person's needs may be used incorrectly as a triage measure to exclude referrals and manage people out of the referral pathway. Services also may develop their own unvalidated cut off criteria for referrals or access to support. There is evidence that such practices are in place; for example, some services have reported using a neurodiversity profiling tool completed jointly by parents and professionals, with the outcome of providing strategies for support rather than referring or triaging people who meet criteria for assessment.

Due to current waiting times and the need for triage, people who qualify for a diagnosis may be missed as criteria for receiving an autism assessment are tightened. Current disparities between services in conversion rates suggest different procedures are being employed that result in different thresholds for assessment. Based on current data, NHS England (2023a) recommends that "to reduce wait times in accordance with national policy commitments, a minimum capacity is needed for at least 1.5 to 2.6% of the population to be referred to an autism assessment service and for at least 1.3 to 2.3% of the population to be assessed for autism". However, it also states that there may be local differences that affect the level of resources needed, so historical rates within an area should be accounted for.

Population screening for autism is not currently recommended (UK National Screening Committee, 2022). Population screening involves inviting everyone from a specified population to undertake a brief test to identify a subset who would benefit from more resource intensive or more invasive tests, or from receiving information or intervention. Targeted screening is when people are screened for a condition because they have been

identified as being at an increased likelihood of having the condition compared with the general population. Targeted screening should be used in autism assessment pathways, so that people are only screened when it has been identified that the person may be autistic. For example, a person (or their family/carer) may believe that they are autistic and approach a clinician asking for support, or a clinician identifies that the person may be autistic.

Autistic people frequently have other co-occurring neurodevelopmental conditions, such as intellectual disabilities, attention deficit hyperactivity disorder (ADHD) and specific learning difficulties (Astle et al., 2022). Many of these conditions also share traits that present in a similar way to traits of autism. In cases where a person exhibits characteristics that may be associated with multiple conditions (for example, sensory reactivity differences), multiphasic screening may be used, where multiple screening measures are completed at once. Screening for an autism assessment in this case may include autism-specific screening measures, more general measures that assess likelihood of having a neurodevelopmental condition, or developmental screening tools.

It is important to note that screening is not always accurate, and the results of screening instruments should always be used alongside clinical judgement. If someone receives a result suggesting they have a low chance of having a condition, this also does not prevent the person from being identified as having the condition later – this is common for autism, where the onset of observable traits may occur at various stages across the lifespan (Wallis & Guthrie, 2024). For screening to be as effective as possible, screening needs to be responsive towards the risk of false positives and false negatives. Tools need to have good sensitivity (that is, the ability of a tool to correctly identify people *with* a condition) with high rates of true positives, and specificity (that is, the ability of a tool to correctly identify people *without* the condition) with high rates of true negatives. This will help to ensure that few autistic people are ruled out of receiving an autism assessment but also that wait times are not exacerbated further by including many people who are not autistic. However, sensitivity and specificity are not the only criteria that tools should be measured against. Guidance from NICE (2017) also states that providers should "be aware that: a positive score on tools to identify an increased likelihood of autism may support a decision to refer but can also be for reasons other than autism; a negative score does not rule out autism." Positive scores on tools may also be indicative of co-occurring conditions, such as ADHD.

False negative or false positive results can be harmful in screening for many health conditions and disorders, including autism. Someone may either be falsely reassured or unnecessarily worried, respectively, by these outcomes, but how this affects screening and assessment for autism may need to be considered differently to screening for other conditions. Failing to detect if a person is autistic may lead to poorer outcomes in education, employment and health by limiting access to support. However, post-diagnostic support is often lacking, especially for adults. Receiving a diagnosis can also have negative repercussions tied to discrimination or a lack of understanding about autism. These outcomes are especially concerning in the case of false positives.

Finally, NHS England (2023a) emphasises the importance of public and provider trust in services, which need to be responsive to the long wait times and provide reliable and evidence-based assessment processes that give all stakeholders confidence in their outcomes. Equality of access underpins operational guidance in place for the delivery of autism assessment (NHS England, 2023b).

Current guidance for appraising tools used for screening and triage

The UK National Screening Committee (NSC), an independent expert advisory group, appraise the evidence that the benefits of screening outweigh the risks and make recommendations to ministers. NSC criteria for population and targeted screening programmes include recommendations for the use of tools for screening purposes. For example, criteria for targeted programmes recommend the use of “a simple test that has evidence of suitable accuracy and technical performance derived from studies in the population in which the test is being used” and that “there should be a diagnostic investigation available for individuals with a positive test screening result, with evidence that subsequent tests can distinguish those who would benefit from interventions from those who would not”. The full list of NSC criteria is included in Appendix 4Appendix 4.

The Joanna Briggs Institute (JBI; Campbell et al., 2015) recommend using the QUADAS (QUality Assessment tool for Diagnostic Accuracy Studies) tool in reviews of diagnostic tests. The tool is used to evaluate the design of studies of diagnostic tests for risk of bias, and applicability. JBI also recommend reporting features of studies including sample demographics; geographical location and setting; persons administering and interpreting tests; and other diagnoses in those without the target condition.

Other criteria also exist for evaluating health measurement tools. The COSMIN (COnsensus-based Standards for the selection of health status Measurement Instruments; Mokkink et al., 2010) checklist is designed to evaluate the methodological quality of studies on measurement properties of tools. This includes consideration of the validity (including cross-cultural validity), reliability and responsiveness.

2. Aims and scope

The key aim of this project was to outline and define criteria that can be used to evaluate standardised tools used for screening and triage for autism assessment pathways in England for children, young people and adults. The aim of setting out these criteria is to support services in making decisions about which tools are most appropriate for use within their service. The project included representation from clinical professionals and researchers, as well as the lived experience views of autistic people and families/carers of autistic people.

Several commonly used tools have been included in this report to serve as examples of how these criteria may be applied in practice. Tools that are in scope for being evaluated using the criteria are those which are being used, or considered for use, in England to support with targeted screening for autism assessments at the referral stage, and for triaging patients within services that conduct autism assessments.

3. Methods

3.1 Stakeholders

We invited stakeholders to join an expert reference group (ERG) and take part in discussions. A total of 19 professionals were involved in this group. This included representation from researchers with expertise in tools used for in screening and/or triage (both autism-specific and general developmental measures), and clinicians involved in instigating referrals and conducting autism assessments.

We also invited members of the Autistica Insight Group to be part of a community involvement group for this project. The Insight Group is made up of around 200 autistic adults and parents/carers of autistic people who have an interest and experience in shaping autism research. We prepared briefing information in plain English explaining the aims of the present project and issued a call to the Insight Group for autistic people who had been diagnosed in adulthood, and parents who had been involved in the diagnostic process for their child (including adults who were supported by a parent). We received 27 expressions of interest and purposely selected nine of these to obtain an even spread of autistic adults and parents and to maximise representation across gender, ethnic/language background and relationship to autism and diagnosis. The group included some people who had reported finding the autism assessment pathway difficult to access. Some members of the community involvement group were also professionals with experience of supporting autistic people through the diagnostic assessment process.

3.2 ERG procedures

The role of the ERG was to:

- identify tools used for screening and triage in the autism assessment pathway,
- identify use cases and referral pathways for tools,
- agree and define appraisal criteria, considering use cases, referral pathways and patient demographics, and
- agree definitions for key terms.

A discussion forum for the ERG, facilitated by Autistica, was launched through Loomio on 25th January 2024, in a closed group. Loomio allows invited users to comment on discussion threads and take part in decisions using polls. ERG members read a briefing document (including a list of well-known tools used for screening and triage for autism assessments, and example screening criteria) and were asked to comment on discussions to select relevant tools for appraisal and recommend appraisal criteria.

The core questions posed to ERG members were:

1. Based on your experience, which tools are commonly used for screening and/or triage in autism assessment pathways?
2. Are you aware of any tools that are being routinely used for screening and/or triage in autism assessment pathways that are not included in the list (defined above)?
3. Do you have concerns about any tools (either in this list or others you are aware of) that may be used for screening and/or triage in autism assessment pathways?
4. Are you aware of any tools that have been adapted to assist with screening and/or triage in people from marginalised groups? (for example, people with an intellectual disability[1], people from ethnic minority backgrounds, and people whose first language is not English)
5. What criteria do you consider to be important to appraise tools for their use in screening and/or triage in NHS autism assessment pathways? These may include general criteria as listed above, but also think about any other criteria you would want

[1] Note that services in England may be called learning disability services, whereas the diagnosis made will be intellectual disability. Throughout this document, we have used the term intellectual disability.

to prioritise when considering autism screening specifically. How might criteria relevant to triage differ from those relevant to screening?

6. How might different use cases, populations or other contextual considerations affect the criteria used to appraise tools used for screening and/or triage purposes?

ERG members were sent two reminders to contribute to discussions before polls opened to select appraisal criteria and tools to review. The forum discussions, as well as existing literature, appraisal criteria and other guidance, were used to generate a shortlist of tools and appraisal criteria. Between 8th and 13th February 2024, ERG members were asked to vote for the criteria and tools they considered most important to this report using Loomio dot-vote polls. These types of polls allocate each voter several dots which can be spread across the different options. For example, a person may choose to place all their dots on a single option they feel to be important, or to spread them across different options. Poll topics were:

- tools for screening – autism-specific,
- tools for screening – general development and co-occurring conditions,
- tools for triage – autism-specific,
- tools for triage – general development and co-occurring conditions,
- appraisal criteria for screening tools, and
- appraisal criteria for triage tools.

A space was available for ERG members to add comments or recommend different options. Some options that received fewer votes than others were combined with higher-ranked options where relevant; for example, application across contexts was incorporated into the umbrella of reliability, and training/expertise needed was included under feasibility. As well as votes in polls, we also reviewed Loomio discussions for other considerations mentioned by ERG members to help evaluate tools. A literature search was conducted on all relevant tools as well as on appraisal criteria relevant to diagnostic screening measures and some of this literature was provided by ERG members in response to discussion topics.

Once the literature review was produced, ERG members were invited to an online workshop to offer any additional insights in terms of relevant literature, or to discuss other gaps or considerations prior to the report's completion. ERG members were asked to help shape the project further by considering how the evaluation of tools against the agreed criteria should be presented to service providers, how service needs should inform criteria, and how communication could be strengthened between service providers and communities.

3.3 Community involvement procedures

The community involvement group were sent copies of six tools (AQ-10, RAADS-14, SRS-2, M-CHAT, SCQ and SDQ – abbreviations expanded in Appendix 2 and Table 2), with a list of questions to consider prior to the focus group session, generated from the ERG discussions and wider literature on acceptability. At the start of the focus group, participants were presented with contextual information on the aims, scope and expected outcomes of the project. Focus group questions and discussions covered views and experiences on screening and triage processes in the autism assessment pathway and views on the tools used. This included questions about their thoughts on such tools being used for screening and triage, their views on the terminology and format of tools and their thoughts about whether anything was missing or could be changed about the tools.

Focus groups were facilitated by three Autistica staff members who all identify as neurodivergent. Focus group members were reimbursed for their time spent on this work.

3.4 Literature review

A narrative literature review was used to appraise the available evidence against criteria agreed by the ERG, focusing on the example tools selected by the ERG. Due to the extensive literature and limited nature of discussions, not all tools which are in use or have been developed could be included in the review. The review prioritised previous literature reviews and meta-analyses of single and multiple tools. Individual studies or findings from grey literature were included where these addressed review gaps (for example, studies not included in a previous review or meta-analysis, studies assessing the value of a general tool as a screening tool for autism, and studies examining differences in performance across demographic groups). Only studies written in English were included, with UK-based studies prioritised for inclusion. ERG members also suggested studies for inclusion in the review. The findings of this review are summarised in Table 5 and Table 6 and outlined in Section 6. The full literature review is included in Appendix 5.

4. Findings

4.1 Criteria selected by the ERG for appraisal of tools used for screening and triage in autism assessment pathways

Providing reasonable justification for the selection of tools used by services is essential for building trust in autism assessment services to deliver best practice. Service providers should use high quality evidence to inform their decisions about which tools to use to assist with screening and triage in autism assessment pathways, as well as how tools are integrated into the process. Providers should be aware of the criteria that are considered important in selection of tools and avoid the use of tools where such evidence is unknown. Using available evidence, providers should weigh the known limitations of tools used for screening and triage against the known benefits. Providers should avoid relying solely on tools that have a limited evidence base for use in screening and triage in autism assessment pathways.

From the list of criteria developed from forum discussions, existing literature, appraisal criteria and other guidance, ERG members used the dot-vote method to select the criteria they felt were most important for appraising tools used, or being considered for use, for screening and triage in autism assessment pathways. Table 1 shows the criteria ranked by ERG members, indicating the percentage of dots allocated to each criterion for the full group.

A full overview of the criteria, including reference to the literature and both ERG and community group discussions and considerations for applying the criteria to tools used in different use case is presented in Section 6.

Table 1. Criteria for appraising tools used for screening and triage, ranked by ERG group members using the dot-vote method. Percentages indicate the percentage of dots allocated to each criterion.

Criteria for tools used in screening	Criteria for tools used in triage
1. 17%: Validity 2. 13%: Acceptability 3. 11%: Reliability 4. 7%: Feasibility and efficiency 5. 7%: Extent of evidence	1. 17%: Validity 2. 14%: Reliability 3. 11%: Extent of evidence 4. 8%: Feasibility and efficiency 5. 8%: Outcomes

6. 7%: Design characteristics of validation studies Lower ranking criteria: Consistency across populations; training/expertise needed (later incorporated into feasibility); flagging of alternative or co-occurring conditions; involvement of autistic community in development (incorporated into acceptability); outcomes	Lower-ranking criteria: Acceptability; design characteristics of validation studies; flagging of alternative co-occurring conditions; consistency across populations

The high ranking of validity and reliability in both categories reinforces the need for tools to be supported by rigorous research evidence. The tools' feasibility for use in real service settings also ranked highly. However, the rankings suggest that tools used in triage may need to be more pragmatic, with more of a focus on outcomes than acceptability.

Criteria should be considered together, rather than in isolation. The importance of these criteria may need to be weighted differently dependent on context or use case. For example, to meet the need for services to be evidence-based, the validity, reliability, extent of evidence and design characteristics of studies are key areas to focus on. There may be some ways services can mitigate for other criteria that are not met that are less dependent on the quality of the evidence.

4.2 Tools selected for appraisal by the ERG

From the shortlist of tools, the dot-vote method was used to gain consensus across ERG members and 17 tools (shown in Table 2) were selected by the ERG for appraisal against the agreed criteria. The criteria are explored in Table 5, Table 6 and Section 6. Most of these tools are questionnaires with self-report and informant versions available depending on the instrument. Many of these tools are freely available, whereas others require a licence to use in practice. Most do not require specialist training. The more general neurodevelopmental assessment tools showed a tendency towards being more suitable for children, whereas the autism-specific tools were more applicable across a range of age groups.

Table 2. Tools selected for appraisal by the ERG.

Autism-specific tools	Population	Administration	Components	Cost
Autism Spectrum Quotient, 10-item (AQ-10)	Adults without an intellectual disability	Self-report	10-item questionnaire	Freely available
Autism Spectrum Quotient, 50-item (AQ-50)	Adults without an intellectual disability	Self-report	50-item questionnaire	Freely available
Ritvo Asperger and Autism Diagnostic Scale-Revised (RAADS-R)	Adults without an intellectual disability	Self-report	80-item questionnaire (25-45 minutes)	Freely available
Modified Checklist for Autism in Toddlers (M-CHAT / M-CHAT R/F)	Children 16-30 months	Parental report	23-item questionnaire Follow-up interview	Freely available Interview is more resource-intensive

Social Communication Questionnaire (SCQ)	Children and adults	Informant report	40-item questionnaire (10-20 minutes)	Licence required
Social Responsiveness Scale 2nd edition (SRS-2)	Children and adults without an intellectual disability	Informant report. Adults can self-report	65-item questionnaire (20-30 minutes)	Licence required
Adult Social Behaviour Questionnaire (ASBQ)	Adults without an intellectual disability	Self-report or informant report	44-item questionnaire	Unclear
Autism Spectrum Screening Questionnaire (ASSQ)	Children and adolescents	Parental report	27-item questionnaire (10-15 mins)	Freely available
Social Attention Communication Surveillance – Revised (SACS-R) ASDetect (app version)	Children up to 24 months	Informant-report and clinical assessment Option for only informant-report in app	12-15 behavioural observations	Training required App is freely available
Neurodevelopmental assessment tools	**Population**	**Administration**	**Components**	**Cost**
Ages and Stages Questionnaire (A&SQ)	Children 1-66 months	Parental report	10-15 minutes	Inexpensive
Developmental And WellBeing Assessment (DAWBA)	Children and adults	Informant and self-report available	Parent interview 50-minutes, child interview 30-minutes, teacher questionnaire 10-15 minutes	Freely available Requires review by clinicians
Children's Communication Checklist (CCC)	Children aged 4-16 years	Informant report	70-item questionnaire, 5-15 minutes	Licence required
Strengths and Difficulties Questionnaire (SDQ)	Children and adults	Informant- and self-report available	25-item questionnaire	Freely available
Early Year's Foundation Stage Profile	Children in first school year	Teacher report	17-item assessment	Routinely used in UK schools
Developmental Behaviour Checklist (DBC)– Autism Screening Algorithm	Children and adolescents 4-18 years	Informant report	29-item questionnaire, 5-10 minutes	Licence required
Conners	Children and adults	Informant- and self-report available – tends to use multi-informant report	10-15 minutes per scale	Licence required
Behavior Rating Inventory of Executive Function (BRIEF)	Children 5-18 years	Informant and self-report	86-item questionnaires, 10-15 minutes	Licence required

4.3 ERG discussion summary

Group discussions highlighted how decisions around tools used for screening and triage are inseparable from the broader systemic issues that make up the contexts in which these tools are used. Issues included:

- concerns about patient confidence in diagnostic pathways and the risk of simplifying screening and triage based on assumptions that all patients are offered the same service,
- how to define clinical need for the purpose of autism diagnosis and care, and differentiate this from social needs,
- impacts of current long waiting times and limited resources on decision-making about thresholds, and how these pressures may risk penalising historically underdiagnosed groups.

ERG members highlighted that the evidence base for using many tools in screening and triage in the autism assessment pathway is limited and often tools are being used in contexts for which they have not been designed nor adequately tested. One researcher stated:

"Several [tools] were designed for use in case-finding and prevalence estimation studies where their unreliability is adjusted for in the study design and analysis, and simply ported into screening without due consideration to the costs and benefits."

In the context of limited NHS resources and growing waiting lists, the group highlighted that use of non-standardised screening tools is showing limited effectiveness at reducing demand on services, while the use of more bespoke tools is associated with a lack of evidence. This may also have implications for whether services can be considered trustworthy and reliable, as they are expected to use evidence-based practice.

ERG members also highlighted a lack of understanding about the true prevalence of autism and expressed concerns that both under- and overdiagnosis make it more difficult to assess how well tools are functioning and how they affect access to services for certain groups. Although there is evidence that some groups have experienced underdiagnosis, such as women and older adults (O'Nions et al., 2023), ERG members agreed that there is a lack of justification for wider population screening due to overstretched services and scarcity of intervention or support available post-diagnosis.

ERG members also raised concerns over AQ-based screening (using all versions of the AQ), despite this tool being in wide use based on NICE recommendations to use the AQ-10 for adults who do not have a moderate or severe intellectual disability. Due to relatively low specificity, it was felt to offer little value as its use resulted in most referrals scoring above the threshold for assessment. Some services also reported using clinical judgement alongside the AQ to identify false negatives. ERG members felt that the well-known nature of the AQ could also impact reliability, because it would be easy to look up online how to obtain a high score for assessment.

There was also concern that some tools lack real-world validity and that the evidence is generally limited to a focus on statistical validity, which assumes the impacts of false positives and false negatives carry the same weight. The group highlighted that less is known about practical consequences in terms of costs and benefits of using tools in clinical contexts, and that screening using these tools also gives little indication of a person's current needs and areas of difficulty. Tools that can capture presentation over a longer period (for example, those that ask about early childhood as well as later life, such as the RAADS-R

and SCQ) may help to assess longer-term needs. Tools also often omit family history, meaning this information may be unavailable at screening and triage despite its value in reaching a clinical judgement. Some ERG members felt levels of need or functioning should be assessed to determine how people are triaged; however, triaging people based on social needs falls outside NHS England guidance, which is based on clinical need. The information needed to triage people based on social need also does not tend to be captured by standardised tools.

ERG members felt that tools that assist with identifying a range of neurodevelopmental conditions could be more valuable for identifying additional clinical need associated with co-occurring conditions. ERG members also pointed out that false positives on screening tools can be indicative of the presence of other neurodevelopmental conditions. One member raised the concern that tools that screen for multiple conditions may be difficult to validate.

ERG members emphasised the importance of ensuring practice is inclusive towards autistic people with an intellectual disability, with some members reporting that some service pathways now include tools which have been adapted or designed to screen those with intellectual disabilities, such as the Diagnostic Behavioural Assessment for ASD (Sappok et al., 2015).

Group members raised concerns that the purpose of using screening tools was not well-understood by patients and their families. They emphasised that trust in the systems was paramount to reduce wait times, which could be achieved by enabling people to feel reassured about outcomes of screening and triage, and the judgement of clinicians. The knowledge and skills of all professionals in the pathway was also considered important for impacting outcomes and trust in screening, triage and assessment. ERG members reported that some services had stopped using standardised tools due to their use of deficit-focused language which some service users found offensive. Group members highlighted that issues around acceptability could diminish confidence in decisions made by services. This may be due to the age of some commonly used tools, or because they were not co-produced with autistic people. One ERG member commented:

“Personally, I don’t know of any commonly used, well validated measures in which the views of the autistic community were an integral part of the development process.”

Another ERG member also suggested that new tools may need to be co-produced to improve acceptability and trust in the tools used.

Discussions highlighted that brief screening tools were considered more feasible for clinicians than options relying heavily on free text. This may not always be considered by patients and informants, who often want to expand upon their answers. One ERG member said:

“I think it is of paramount importance that the patients’ and carers’ views are listened to and taken on board.”

This may be harder to achieve at the point of triage, where a tool may be used in isolation to ascertain if a person meets the threshold for screening. ERG members pointed out that longer, clinician-assessed tools lack scalability for implementation as screening tools compared to brief patient- or informant-rated measures but might have applications for triage. More evidence is needed for their use in this context. One member suggested that training clinicians such as health visitors to administer screening tools may increase awareness of autism.

Attention to different contexts is also important, as one practitioner reported that most of the referrals that they received were from mental health services and nearly all met screening thresholds. This suggests that screening may not be as necessary for referrals received from clinicians with high levels of existing expertise. People may also be more likely to meet screening thresholds when referred from mental health services as the prevalence of certain mental health conditions in autistic people is higher than in the general population (Lai et al., 2019). Some common symptoms of mental health conditions may also overlap with symptoms of autism, leading to false positives on screening measures.

4.4 Community focus group discussion summary

Community involvement group members discussed their views and experiences of screening and triage processes and tools. Discussions focused on acceptability as an area with less evidence from published research. This may be an important factor in informing a person's confidence to accept the decision of a clinician, which impacts on demand on pathways through the need for re-assessment when people seek a second opinion. Group members reported barriers to completing questionnaires for screening and triage in the autism assessment pathway. These included ambiguous questions, unspecified context and a lack of ability to capture fluctuations in their presentation. One group member provided an example of how answers to questions depend on the context:

"Here's one: 'I feel self-confident when interacting with others.' [...] As [a professional]? Yes. I was self-confident because I knew what I was doing [at work] and they didn't."

Group members reported that some questionnaires included items that they felt were outdated or culturally biased. They felt the questions should reflect both current diagnostic criteria and lived experience, such as assessing co-occurring conditions or other markers reported as common by autistic people (for example, digestive issues, depression and motor difficulties). Group members noted that the phrasing of questions was often considered negative, value-laden or deficit-focused. There was less awareness towards atypical presentations, those who surpass developmental milestones, masking of autistic traits, and those who present as more sociable and empathetic.

"I believe they should screen for both delays as well as being ahead of development to fully embrace what makes this condition a spectrum."

"Unless I paint the very worst picture of the worst time that my child might have had, I won't get any support […] it's awful having to do it. It's very triggering and horrible."

Some group members reported that completing the questionnaires was sometimes a positive and validating experience. They emphasised that taking a family history was important, but most tools do not capture this information. Group members recognised the value that informants can provide to the assessment processes, but identified barriers, including not having an informant available, informants not wanting to engage due to stigma, and family members who may have a lack of insight into how their family norms may differ from those of other families.

The group felt that a choice between paper and digital versions of questionnaires should be offered. Some group members reported that digital versions could reduce overwhelm by presenting questions one at a time and assist those who may struggle to return a paper questionnaire back to a service, while others felt completing questions on paper allowed

them to take their time. Services requiring telephone communication was reported as a barrier to accessing a referral.

The group also discussed the impacts of being precluded from assessment based on screening outcomes in services commissioned and provided by the NHS, including:

- the use of the private sector, with some reporting taking out loans to pursue a diagnosis,
- increased reliance on other services such as NHS Talking Therapies, which may be less appropriate for the needs of autistic people,
- feeling shunned or disbelieved, and
- not receiving support on time.

A lack of understanding of autism in female, non-binary and transgender people, and in people from ethnic minorities was felt to contribute to these impacts. Group members felt that tools were not always sensitive to differences across groups.

"Women of colour are often overlooked as there is not enough representation and they are often considered to have attitude issues, etc. My wife was turned away from seeking diagnoses despite having multiple family members with [an] autism diagnosis and she was not taken seriously. She had to pay for her screening privately to be taken seriously."

Several group members felt that involving the autistic community in the design and delivery of assessment services would increase acceptability. They also felt that triage processes should focus more on individual circumstances and support needs, rather than scores on questionnaires (for example, if someone is in crisis, applying for benefits, sitting exams, or needs accommodations at work). Some group members reported experiences of triage operating in this way, including those who were service providers themselves. These issues contrast with commissioning guidelines highlighted by ERG members that require screening and triage to focus on clinical need, instead of social needs. Community involvement group members also reported differences in experiences between geographical areas, suggesting that services operate according to a 'postcode lottery' regardless of clinical guidelines. Group members who had moved area while on the waiting list reported having had to begin the process again. Others reported being told by health professionals that there was not an autism assessment pathway in their area.

"They said the information from [previous clinical commissioning group (CCG)] was too old (even though autism is lifelong!) so the CCG refused the referral."

"We have to make sure people can access the screening tests as a first step. [My GP] didn't deny I would be autistic; she just didn't think there was a pathway."

Group members felt that a lack of understanding about reasonable adjustments and pre-diagnostic support in settings such as schools may push more people down the assessment route when only simple accommodations may be needed.

"My niece […] can't be referred for assessment because there's no funding, while her school is desperate to get her assessed so they can give her the support she needs to enable her to go to nursery"

The issues reported by group members imply that supporting people to access resources and information prior to or during the wait for an autism assessment may reduce the impacts of waiting times and outcomes of screening and assessment processes for patients, families and services. Social prescribers may be well-placed to support people to access resources in their communities at any point throughout the assessment pathway and are a

recommended referral pathway for those who may not meet criteria for an autism assessment (NHS England, 2023b). Environments such as schools and workplaces could make universal adjustments and work towards strengthening inclusive practices without requiring people to have a formal diagnosis to access support. Offering increased support that does not rely on a diagnosis could also reduce demand on autism assessment services and adjacent services.

5. Use cases

Tools may require a different combination of appraisal criteria or different thresholds for appraisal dependent on use case. Considerations for usage may also include the reporter (for example, self- versus informant-report) or demographics, such as age or gender. Tools used for screening and triage in autism assessment pathways may be used in different cases, including for the purpose of:

- identification of possible autism by a GP or SENCO, and subsequent referral supported by a screening instrument,
- screening and/or triage, to be completed by healthcare professionals within a community mental health team, or any team that is not a specialist autism team, and
- triage completed by healthcare professionals within specific autism assessment services.

Tools can show variations in performance across different use cases due to differences in patient population, who completes the tool being used to support screening and/or triage, the expertise of the person completing the tool, and the feasibility of using different tools (for example, if only brief consultations are conducted in a particular setting).

Tools may also perform differently based on individual characteristics, such as the age, gender and ethnic or cultural background of a person, and the presence of co-occurring conditions. Where tools have been designed with specific groups in mind, they should be used in these groups, although this is not always possible. Due to potential biases in the design and validation, some tools may have not be validated across different genders or ethnicities. Some tools may mitigate this, for example, by using different scoring procedures according to the person's gender. However, studies examining gender differences often only categorise people as male and female, providing no insight into the applicability of these tools for transgender and non-binary people.

Extra care may also need to be taken when using tools in populations with additional or complex needs including intellectual disabilities, Down syndrome, sensory impairment, mental health conditions, and behavioural disorders, where tools have not been validated for these populations.

Recommendations may also differ depending on whether tools are intended to be used for screening or triage (or screening and triage) and whether they are used in combination or multiphasic screening approaches. There is currently a lack of evidence available considering the use of tools for triage in the autism assessment pathway compared to their use in screening. See Table 5, Table 6 and Appendix 5 for an overview of the available evidence to support the use of tools for screening and/or triage in the autism assessment pathway.

6. Criteria for evaluating tools used for screening and triage for autism assessment pathways

This section contains the full list of criteria that can be used to assess the suitability of a tool for screening and/or triage in the autism assessment pathway, as agreed upon by the ERG. Each subsection will define the criteria, demonstrate how they apply across use cases, and provide examples from research. The research examples refer to the 17 tools selected based on ERG consensus (nine tools designed specifically for autism screening, identification or profiling and the eight tools that measure generic developmental differences or co-occurring conditions). The full list of tools is presented in Table 1. Each subsection also contains some self-assessment questions to guide decision making around which tools best fit the criteria. The full list of self-assessment questions is provided in Appendix 6. Detailed information about the literature relating to the criteria and tools is available in Table 5, Table 6, and Appendix 5.

The criteria identified by the ERG are as follows:

- Validity and reliability
- Acceptability
- Feasibility and efficiency
- Extent of evidence
- Design characteristics of studies
- Outcomes
- Flagging of alternative/co-occurring conditions
- Consistency across populations

6.1 Validity and Reliability

The validity of a tool refers to whether the tool measures what it intends to measure. The reliability of a tool refers to whether the output is consistent when repeated. Indicators of validity include sensitivity (the ability of a tool to correctly identify people *with* a condition) and specificity (the ability of a tool to correctly identify people *without* the condition). These constructs are valued between 0 and 1; a sensitivity value of 1 indicates that 100% of people identified as autistic are true positives, and a specificity value of 1 indicates that 100% of people identified as not autistic are true negatives. To provide a measure of accuracy, these values can be plotted against each other using a method called ROC curve analysis. This produces a measure of accuracy called the Area Under Curve (AUC). An AUC figure of 1 indicates a perfect test whereas 0.5 indicates a test which cannot distinguish between people with and without a condition. Specificity, sensitivity and AUC are included for each tool in the literature review in Appendix 5, where these figures are available in the literature. No screening tools for autism are infallible, and as the employment of the tool is not intended to confirm a diagnosis, service providers will need to accept an inevitable margin of error but should consider the impact of rates of false positives and false negatives on service delivery, such as waiting lists and re-referrals.

Other aspects of validity include the tool's ability to predict the diagnostic outcome (predictive validity), its ability to discriminate between autism and other conditions (discriminant validity), and whether a tool is measuring what it claims to measure (construct validity). Tools should be evaluated in a context like that in which they will be used

(ecological validity). Skewed results, such as floor or ceiling effects (that is, where everyone scores very high or very low on the measure), may indicate issues with the validity of a tool.

Reliability includes inter-rater reliability, which is the ability for the ratings of different respondents to result in similar scores (for example, self- and informant- report, or the ratings of two different clinicians); test-retest reliability (repeating the test should result in a similar score) and internal consistency (a value demonstrating whether items in the scale are measuring the same construct). The tool should function the same way across contexts.

6.1.1 Use cases

Whilst sensitivity and specificity of tools may be high in some research studies, it is important to consider the methods used in the research, with particular focus on the population and setting. How representative are these studies of the clinical setting in which the tool will be used, the population that the service is likely to cover, and the referral mechanisms used in the service? For example, consider that some autism assessment tools have been tested for the purpose of identifying autistic traits and cases of possible autism in general population samples. This may not be representative of the populations presenting at screening and triage, who may present with elevated autistic traits compared to the general population, whether or not they meet the criteria for a diagnosis of autism. Someone who has been identified as being possibly autistic may also have other co-occurring conditions or symptoms of co-occurring conditions that are less likely to be found in general population samples.

The importance of different types of reliability may also differ dependent on use case. Inter-rater reliability will be important if the screening or triage process uses the same tool to access views of multiple informants. Test-retest reliability is also important to understand the consistency of a tool's performance more generally, however, this will be highly relevant for services where a test may be repeated (for example, at both screening and triage stage).

It is also important to consider that reliability may be reduced when respondents are overly familiar with tools, such as when tools are well-known and easily accessible (such as being freely available online).

6.1.2 Examples from research

Reviews of the M-CHAT show variations in performance across settings, with evidence of poorer performance in clinical settings than research settings (Levy et al., 2020; Wallis & Guthrie, 2024).

There has been extensive research on the A&SQ for children up to age 5, with limited studies suggesting it performs well at identifying autism, with high agreement with M-CHAT scores.

The AQ is widely used for self-reported methods of screening in adults. Some previous reviewers have suggested this tool is unsuitable for use in screening and triage, with unsatisfactory validity leading to high numbers of respondents scoring above the threshold (Baghdadli et al., 2017; Brugha et al., 2020; Wigham et al., 2019). AQ scores have also been found to correlate poorly with diagnosis (Ashwood et al., 2016; Conner et al., 2019; Jones et al., 2021).

The ASBQ can be completed by an adult or an informant. Data on its effectiveness are limited but suggest satisfactory validity and reliability (Baghdadli et al., 2017; Horwitz et al., 2016).

The SDQ was found to perform consistently across settings but lacked data on validity across diverse demographic groups (Warnick et al., 2008). Some tools which rely on multiple informants, such as the SDQ and SCQ, have also shown discrepancies in ratings depending on the person completing the questionnaire (Hollocks et al., 2019; Findon et al., 2016).

The SCQ was found to have higher predictive validity than other comparable measures, but its validity was lower in clinical samples (Charman et al., 2007; Metcalfe et al., 2020).

Findings on the SRS-2 are mixed and suggest sensitivity and specificity of this measure is greater in children than in adults (Bruni, 2014).

Self-assessment questions – Validity and reliability

- What was the original purpose of the tool(s) in question?
- Has there been research that has carried out psychometric validation of the tool(s)?
- What aspects of validity and reliability have been studied? What is missing?
- Have the findings been peer-reviewed?
- Has the tool been tested in settings and populations that are representative of the service/setting?
- How much exposure have the population had to this tool already? Will prior knowledge affect its reliability?

6.2 Acceptability

Acceptability concerns how well a tool is received by patients, their families/carers, assessment teams and referrers. Indicators of a tool's acceptability might include how the tool is presented, how burdensome it is to complete, relevance to the needs of the users, and any difficulties completing the tool.

Acceptability has been captured to a lesser extent within research literature, which tends to focus more on psychometric properties of tools. Patients and clinicians may need to be approached about their experiences to gauge acceptability.

ERG members highlighted that in some cases, bespoke and unvalidated tools were used because they had higher acceptability to patients than standardised tools. Where unevidenced or bespoke tools are being used because they are felt to have higher acceptability for patients, it is important to also consider how these tools perform on the other criteria identified and agreed by the ERG.

During the community involvement focus group, community members responded to questions on aspects of tools they found acceptable and any issues affecting acceptability. Example quotes relating to these aspects are presented in Table 3. Some of these aspects overlap with other criteria, such as feasibility (for example, "Quick to complete") and consistency across populations (for example, "Male-oriented").

6.2.1 Use cases

Acceptability of tools may vary dependent on the population and setting. For example, in some services, acceptability to the community may be impacted by the tool's availability in different languages.

Table 3: Examples of community involvement group perspectives on more and less acceptable features of screening tools.

Features of tools viewed as more acceptable	Features of tools viewed as less acceptable
“Quick to complete”	“Ambiguity”
“Easy to understand”	“Insensitive wording”
“Room for context”	“Male-oriented”
“Easy-read format”	“No neutral answer”
“Relevant”	“Assume insight”
“Examples included”	“Outdated”

6.2.2 Examples from research

The literature revealed that there was little consideration towards acceptability and trust in services in the design and evaluation of standardised tools, although some research does explore the acceptability of existing tools. Tools being returned incomplete may be indicative of their acceptability

Respondents on the RAADS-R reported difficulties with ambiguous questions and rigid response options, and value-laden language perceived as stigmatising (Stacey & Cage, 2022). Two of the response options on the RAADS-R may also be redundant in terms of its performance, suggesting this tool could be simplified for greater acceptability (Sturm et al., 2024).

Scores on the SRS-2 differed according to whether questions referred to an in-group or out-group in relation to the respondent. This demonstrates issues with how items may be interpreted (Gernsbacher et al., 2017).

Some older tools may use outdated language; however, an older tool is also more likely to have an established evidence base. For example, a screening tool suitable for people with an intellectual disability and used as part of assessment in current practice is the “Scale of pervasive developmental disorder in mentally retarded persons” (Kraijer & de Bilt, 2005). The tool has shown good sensitivity, and it is available in translated versions, potentially making it more inclusive (Metcalfe et al., 2020). However, it is unlikely the name of this tool would be considered acceptable in a modern context among autistic people, families/carers and clinicians.

Self-assessment questions – Acceptability

- Is there any research available that explores the acceptability of the tools for different user groups?
- Does its acceptability to patients align with its acceptability to professionals involved in screening and triage in the autism assessment pathway?
- How burdensome is the tool to use – for patients, for professionals and for other users?
- What difficulties often arise when completing or scoring the tool?
- What aspects of the tool might indicate its acceptability for the different user groups? How can this be assessed?
- Are there ways to mitigate for low levels of acceptability for a tool that is otherwise well-evidenced? For example, would it help to give clearer information about why certain terminology is used?

6.3 Feasibility and efficiency

The feasibility of a tool refers to whether it is practical and viable to use for a specific purpose. Efficiency refers to whether a tool is effective at its intended aims.

For a tool to be feasible to use in services, it should be cost-effective and represent value for money. Some of the tools that may be used for screening and triage in the autism assessment pathway are free to access, while others require a licence for use. Paying for licenced tools may be more valuable if the tool offers better performance overall than the freely available alternatives.

Validity and reliability influence feasibility and efficiency. Providers will need to consider the impacts of false positives and false negatives on the operation of their service, as well as the impacts of these on adjacent services. Not using a tool for its intended purpose may have implications for costs and efficiency of services, as the tool may lack sensitivity and specificity in that context.

Other factors that will influence cost-effectiveness include how much clinician time and resource is needed to administer or score a tool, and whether any training is needed to do so. A minority of studies have carried out cost-benefit analyses of tools based upon some of these criteria.

It is also possible that patients may not complete tools in their intended way. For example, they may want to add detailed written responses to a tool that uses a Likert scale. This may be an ineffective use of time for patients, who may have this extra detail overlooked when the tool is scored, or for clinicians if they take the time to read detailed answers.

6.3.1 Use cases

Settings such as GP practices may require a brief consultation with a clinician who is not an expert on their patient, whereas secondary mental health services may involve clinicians who work one-to-one with a patient over a longer period. These considerations may affect which tools are possible to use and how much information can be conveyed in a referral.

In geographical areas where a high proportion of the people accessing a service do not have English as a first language, a tool may need to be translated. It is important to note that not all translated versions of tools will be validated. A tool may also need to be presented in an Easy Read format for people with an intellectual disability.

Informant-report is usually a vital part of autism assessment to assess a patient's developmental history, and some tools used for screening and triage in autism assessment pathways are designed to be completed by informants who knew the person as a child. However, this may not be feasible for older adults who are less likely to have access to a person who knew them as a child.

6.3.2 Examples from research

An advantage to some of the generic tools is that these are used routinely in education or healthcare. An example is the Early Year's Foundation Stage Profile, which has high feasibility to support with screening of primary school-age children because this makes use of data routinely collected by schools. It has also been found to correlate with autism diagnoses (Wright et al., 2019). The two-stage version of the M-CHAT, which includes a clinical interview and a parent-report questionnaire, has been shown to be more reliable (Charman & Gotham, 2013), but will also be more resource-intensive. Note how these examples indicate the need to balance the quality of the tool with its feasibility for use.

Wallis and Guthrie (2024) have reported that where some referral pathways have a very high rate of referrals with scores above the threshold (such as referrals from mental health services), services in the United States have been known to skip screening processes and conduct universal diagnostic assessments on all referred patients.

Ashwood et al. (2016) analysed the diagnostic accuracy of the AQ and found that "for every 10 patients denied a referral, £23,050 would have been saved but seven [autism] cases would go undiagnosed" (Ashwood et al., 2016, p.2601). This may have implications for adjacent services, such as mental health care and general practice, as well as influencing the acceptability of using this tool.

In a sample of children with cerebral palsy being screened for autism, there were lower completion rates for instruments including the ASSQ and SDQ among parents of those with the highest levels of complexity, demonstrating that screening procedures were not feasible for this group (Påhlman et al., 2020).

The AQ and RAADS-R are widely used for self-reported methods of screening in adults. These tools have high feasibility as they are easily available and straightforward to use and score, though this may lead to prior familiarity with these tools, which can skew results.

Self-assessment questions – Feasibility and efficiency

- Has the cost-effectiveness of the tool(s) in question been evaluated in the appropriate context? How long ago was cost-effectiveness evaluated?
- Is there any evidence that a tool may be more effectively used in combination with others?
- How much clinician time is required to use the tool? Is clinician time required to support patients with completing tools at referral? How much clinician time is required to review a tool following a referral, or for triage?
- Does the tool require a licence to use?
- Do clinicians have the appropriate technology and training to support using the tool in its intended way?
- Can communication with service users be improved to ensure tools are completed in their intended way (such as explaining why a tool is scored a certain way)?
- Are tools presented in accessible formats?

6.4 Extent of evidence

While a tool may be shown to perform well in some studies, it is important to consider the whole evidence base and whether findings have been consistently replicated by multiple studies. Reviews and meta-analyses bring findings together to assess pooled outcomes and appraise the quality of literature overall. These types of research may give an indication of the total population of samples across all studies and cross-examine features of studies including potential biases and design characteristics (see Section 6.5). Reviews themselves will also need to be high-quality and free of conflicts of interest.

Publication bias, which includes the tendency to over-report positive results and not publish negative results, may also influence what is captured in the evidence base. Consider looking to grey (non-academic or unpublished) literature to supplement findings from published scientific literature.

6.4.1 Use cases

Where possible, it is important to consider research that is appropriate to the use case. There is likely to be less evidence that has evaluated tools when applied in specialist settings and populations such as older adults' mental health services. There is also likely to be more evidence around the use of tools for screening than for triage. Other criteria may need to be examined when deciding on which tools work best for triage.

6.4.2 Examples from research

The AQ and SCQ appear to have been evaluated more than the other tools included for selection. Studies have frequently used biased methods of patient selection, study design and choices of index tests. Most research has focused on the use of standardised tools as screening instruments for identifying the need for assessment, rather than for triage.

There has been a lack of attention in research towards the most suitable tools for screening and triage of older adults in autism assessment pathways (Summerill et al., 2024). There is also a lack of peer-reviewed, published evidence for some of the tools that assess broader needs (and are currently in use in some services to support screening and/or triage), such as the Do-It Profiler and the Neurodiversity Profiling Tool. This shows why acceptability should not be the only criterion on which tools are appraised when aiming to use evidence-based practice to build trust in services. When designing tools for acceptability, their psychometric properties and feasibility for use in services should be rigorously assessed before being applied in practice.

Self-assessment questions – Extent of evidence

- Has the tool been included in any meta-analyses or systematic reviews? If so, how many reviews, and how many studies of the tool have been identified?
- How long ago were meta-analyses/reviews carried out? Is there additional research that has been produced since?
- Are findings on the tool consistent or do they vary? On what aspects do they vary?
- What do reviews and individual studies indicate about the use cases the tools have been applied to?
- Who has carried out studies on the tool? (for example, if all evaluations come from the team who designed the tool, there is a high risk of bias)

6.5 Design characteristics of studies

It is important to appraise the design characteristics of studies to determine the risk of unintentional or intentional bias in how tools have been evaluated. A simple way to do this is to examine systematic reviews relevant to tools of interest, where these are available.

There may be more recent studies missed by reviews or some tools may not have been reviewed at all. If available, reading the primary research should help to identify potential issues with study design, provided reporting is accurate and clear. Issues that may arise include sampling methods (such as small sample sizes), stringent inclusion and exclusion criteria that may have excluded groups who are most likely to be using the tool in clinical settings, blinding of researchers to participant allocation, the reference tests used to compare the performance of the tool, and loss to follow-up.

6.5.1 Use cases

Attention should be paid towards the samples included in research and how well these represent the population within the intended setting for use of tools.

6.5.2 Examples from research

Bagdhadli et al. (2017) found evidence of bias across adult autism screening tools due to issues such as sampling techniques and the reference tests used. Overall, most of the studies testing the psychometric properties of these tools did not apply their methods in naturalistic settings that would mirror current NHS screening and triage procedures, or current pressures facing service decision-makers. Indeed, Jones et al. (2021) evaluated RAADS-R for screening in the setting of an NHS specialist autism service and found that it was ineffective in this setting.

Self-assessment questions – Design characteristics of studies

How do previous studies on the tool(s) compare on:

- Methods used to verify diagnosis (for example, self-report of an existing diagnosis; diagnostic assessment by a clinician; use of another tool as a reference test)?
- Use of reference tests – Which tools have been used as the reference test? Are they well-validated? Were all participants given the same tests? Were reference and index tests rated without knowledge of results on the other test?
- Sampling – What size samples have been used? How were samples recruited? Was any randomisation used? Do the inclusion/exclusion criteria represent people using services?
- Design of procedures – Have case-control designs been avoided (that is, comparing scores of autistic people with scores of non-autistic people, which would be required to determine whether the tool is successful at identifying autism)? Has there been follow-up to assess any change over time?

6.6 Outcomes

Tools may be associated with different outcomes, including conversion rate (number of referrals that result in diagnosis), the outputs they produce (such as, a single score, a cumulative score from subscales, or a profiling tool). The type of output may have different applications in informing next steps for assessment or support.

6.6.1 Use cases

Some tools may not capture information that contributes to clinical judgement on whether to make a referral or how to triage a patient. For example, many tools do not capture family history, a person's developmental history, or their current risk level. The expertise of clinicians within the setting may influence the amount of information that needs to be gathered and how to mitigate against the limitations of tools.

6.6.2 Examples from research

There is little research available that examines which tools commonly lead to specific outcomes for triage in the autism assessment pathway. The five-factor structure of the SDQ was found to be a poor fit for understanding the profiles of autistic adolescents. Apps such

as ASDetect may help parents of autistic children to assess their child's needs without input from a clinician (Barbaro & Yaari, 2020; Green et al., 2022). The DAWBA showed mixed results as a screening tool, though this is not its intended purpose, and performance was best when combined with an ADOS assessment (McEwen et al., 2016).

Self-assessment questions - Outcomes

- What type of output does the tool produce? Who is it useful for? How does it help to inform next steps?
- What information is not captured by the tool that may be crucial to guiding clinical judgment?

6.7 Flagging of alternative or co-occurring conditions

Many neurodevelopmental conditions, including ADHD, and intellectual disabilities, co-occur more frequently with autism compared to their prevalence in non-autistic people. Furthermore, assessments may identify that, for some people, their needs could be better explained by a condition other than autism; false positives on screening or a score just below the threshold may be a flag that another condition could be present. However, determining whether a screening result is a false positive often requires further assessment. There may be benefits to tools which can screen for alternative or co-occurring diagnoses, including reducing waiting times for patients who need to be referred onto other assessment services, and reducing burden on services by redirecting patients to more appropriate pathways.

There are several tools identified in this report that assess wider neurodevelopmental concerns and have been previously validated. Although their evidence in screening for autism is relatively weak compared to autism-specific tools, they may help to indicate if someone has a need other than autism or a potential co-occurring condition that may increase clinical need.

6.7.1 Use cases

The referral setting may affect the population being screened. For example, referrals from mental health services may include more people with co-occurring conditions than screening in a general population sample (the latter is used more in research). Using tools that assess for co-occurring conditions alongside autism-specific tools may support with screening and triage decisions for referrals that come from these services.

6.7.2 Examples from research

Some tools, including the AQ and SRS-2, may be less effective at distinguishing autism from other conditions such as generalised anxiety, dementia and psychosis. For example, the AQ showed a high rate of potential false positives in populations with generalised anxiety disorder (Ashwood et al., 2016). Such evidence may explain why some clinicians in the ERG reported that referrals from some mental health services had a high rate of people meeting the threshold for autism assessment, making the justification for screening tools less apparent in referrals from this setting. This may be due to the tool being used lacking specificity in clinical mental health populations. A study on the ASBQ found that it successfully discriminated between autism and mental health conditions, including schizophrenia and depression (Horwitz et al., 2016); however, the extent of research on this tool is limited.

Self-assessment questions – Flagging of alternative or co-occurring conditions

- What co-occurring conditions need to be identified or ruled out in the population in question?
- What does research suggest about how well the tool can differentiate possible autism from other conditions?
- What does research suggest about how well the tool can indicate the presence of conditions that co-occur with autism? How does this affect triage, outcomes and support?
- If the tool assesses broader neurodevelopmental traits (that is, it is not specific to autism; for example, the SDQ or CCC), has it been validated for screening or triage in autism assessment?

6.8 Consistency across populations

To ensure processes of screening and triage are fair, inclusive and trustworthy, tools should demonstrate good performance against the above criteria across different subgroups including gender, age, ethnicity and disability. A range of tools may need to be recommended where performance differs across groups.

6.8.1 Use cases

Different areas have different demographic profiles, which may need to be accounted for in determining how to make a service accessible and inclusive. Recent research suggests autistic women and girls may have historically been overlooked (O'Nions et al., 2023; Russell et al., 2022) and that being gender-diverse is associated with later diagnosis in autistic people (McQuaid et al., 2024). Tools have often been developed and tested on predominantly cisgender male samples, which may contribute to this bias.

6.8.2 Examples from research

Some tools, such as RAADS-R and SCQ, show differences in performance according to age and gender (Sturm et al., 2024; Wigham et al., 2019; Charman et al., 2007), while the AQ demonstrates gender bias between items (Belcher et al., 2023). The BRIEF was reported as being adjusted for gender (Walker & D'Amato, 2006). The SRS-2 has been found to have greater validity in children than in adults (Bruni, 2014).

The AQ shows evidence that cultural values may affect interpretation of items (Genovesi et al., 2023). Research on the BRIEF demonstrated no significant differences in performance by ethnicity (Walker & D'Amato, 2006). Some tools, such as the M-CHAT, have been translated and validated in multiple languages, which may improve its applicability across cultures; however, cultural bias was found in diagnostic processes and outcomes more generally even where this tool was in use (Wallis & Guthrie, 2024).

Some tools such as the Diagnostic Behavioural Assessment for ASD have been designed specifically for screening for autism in people with an intellectual disability (Sappok et al., 2015). This tool has been found to be less suitable in those with more severe intellectual disabilities (Metcalfe et al., 2020), so the exact population will determine its utility. Other tools such as the CCC show more consistent performance across people with and without an intellectual disability, and for different types of intellectual disabilities (Charman et al., 2007).

The M-CHAT, SRS-2 and SCQ have lower specificity in children with intellectual disabilities and behavioural problems (Charman et al., 2007; Metcalfe et al., 2020).

One study found that combining the M-CHAT and SCQ was less effective in children with additional sensory impairments compared to those without (Metcalfe et al., 2020).

Self-assessment questions – Consistency across populations

- Has previous research investigated whether the tool's performance differs by age, gender, ethnicity and culture or disability?
- What aspects of the service's screening and triage procedures affect access or outcomes for people from marginalised or underserved groups? What is the service doing to ensure processes are fair and inclusive? Does this have any influence on which tools or procedures can be used?

7. Summary of the literature

7.1 Evidence summary tables

Table 5 and Table 6 summarise the findings of research included in the narrative literature review (see Appendix 5), and rate each of the chosen example tools against the criteria agreed by the ERG using a Red, Amber and Green rating system. Table 5 includes ratings for the nine autism-specific tools selected by the ERG and Table 6 includes ratings for the eight general neurodevelopmental assessment tools. Table 4 provides a key to interpreting Tables 5 and 6.

The tables are not intended to dictate to services which tools they should use in an absolute fashion. Rather, they are intended to highlight how to apply criteria in decision-making about which tools are appropriate for use and show the potential benefits and risks of these tools as identified in research evidence. The summary tables may not account for nuance that is captured in the original literature, nor do they cover all available previous research. The tables cover a limited selection of studies that were the most relevant to the report. The extent of the evidence and characteristics of research reviewed is important to bear in mind. For example, a green rating for validity may arise from a small number of biased studies which have found high validity and reliability values, but this may have occurred because there is a lack of published research that has disproved these values or that has been conducted in a robust manner. The colours in the table will appear paler where there is less evidence for a tool being used in the context of screening and triage in the autism assessment pathway.

The findings shown in the table may not consider the consensus of clinicians, commissioners or other stakeholders using these tools in practice. The populations against which these tools have been evaluated are often unrepresentative of those within NHS autism assessment pathways. Different service contexts and processes may inform how services decide to use the information in this table. For example, some service providers may choose to use a multiphasic approach by selecting several tools or decide to use different tools for screening and triage based on the criteria considered important at these stages of decision-making.

Table 4: Key to tables 5 and 6.

Criteria	Evidence reviewed (N = number of studies included in paper)	Research design characteristics	Validity and reliability	Feasibility and acceptability	Consistency across populations/co-occurring conditions/use cases
Findings extracted from research	Extent of evidence identified for review, **relating to use as a screening and/or triage tool for autism assessment pathways** Paler colours indicate less evidence for the reported findings.	Risk of bias – quality of research design **assessed by previous scientific reviews**	Values from previous research indicating sensitivity, specificity, AUC (a measure of test accuracy) analysis, predictive value, construct validity, concurrent validity, internal consistency, inter-rater reliability and test-retest reliability	Findings from previous research on costs, time taken, training required, ease of use, and acceptability to community and clinicians	Findings from previous research on ability of tools to identify co-occurring conditions, differentiate between alternative potential diagnoses (for example, to distinguish autism from ADHD), and to show consistent results across demographic groups
Green ratings		Green = low risk of bias identified	Green = Studies indicate high validity and reliability in for use as a screening and/or triage tool in the autism assessment pathway	Green = Studies consistently indicate tools meet feasibility and acceptability criteria	Green = Studies indicate largely positive findings relating to above criteria
Amber ratings		Amber = moderate/mixed evidence of bias	Amber = Moderate values or mixed findings	Amber = Studies suggest some concerns about ability of tools to meet feasibility and acceptability criteria	Amber = Studies show moderate or mixed ability of tool to meet above criteria
Red ratings		Red = high risk of bias identified	Red = Low values or limited evidence	Red = Studies suggest tools fall short of feasibility and acceptability criteria	Red = Studies indicate tool shows show poor performance to meet above criteria
No colour		Blank = insufficient evidence	Blank = Insufficient evidence on validity/reliability in autism screening/triage	Blank = Insufficient evidence addressing feasibility and acceptability for screening/triage	Blank = Insufficient evidence for performance against these criteria

Table 5: Evidence for use of autism-specific tools for screening and triage in the autism assessment pathway.

Autism-specific tools	**Evidence for use in autism screening** (N = number of studies included in paper)	**Research design characteristics**	**Validity and reliability**	**Feasibility and acceptability**	**Consistency across populations/co-occurring conditions/use cases**
AQ-10 AQ-50	Baghdadli et al. (2017; N=22) Summerill et al. (2024; N=6) Wigham et al. (2019; N=5) Belcher et al. (2023; N=1) Conner et al. (2019; N=1) Murray et al. (2016; N=1) Brugha et al. (2020; N=1) Ashwood et al. (2016; N=1) Genovesi et al. (2023; N=1)	Moderate risk of bias from samples and comparison tests used	Mixed findings for sensitivity and specificity – AQ-50 may have higher sensitivity but shows wide variation Other aspects of validity show mixed findings Moderately correlated with predicting ADOS outcomes Good reliability overall (better for AQ-50)	Easier to complete than RAADS-R Widely available online No training required Screening accuracy (AQ-10) may continue to be affected by incorrect NICE guidance issued (revised in 2021) Cultural norms affect interpretation of questions	The AQ potentially shows bias by gender (results vary) Caution should be applied in people with generalised anxiety disorder, who have been shown to score higher on the AQ, leading to more false positives in screens from this population There is a lack of evidence for its performance in screening autism traits in older adults
RAADS-R	Baghdadli et al. (2017; N=3) Wigham et al. (2019; N=1) Brugha et al. (2020; N=1) Sturm et al. (2024; N=1) Stacey & Cage (2022; N=1) Jones et al. (2021; N=1)	Moderate risk of bias from samples and comparison tests used	Fair, but limited evidence for sensitivity and specificity Other aspects of validity show mixed findings Poor ability to predict diagnostic outcome. Satisfactory reliability	Respondents reported difficulties with ambiguous questions and value-laden language One study found a lack of justification for having four response options and using an ordinal scale, suggesting the tool may be overly complex Ability to capture current and past presentation may be helpful for understanding patient history Readily available with no training required	Shows differences in psychometric performance by age, gender and diagnosis

M-CHAT/ M-CHAT-R/F	Levy et al. (2020; N=19) Santos et al. (2024; N=6) Charman & Gotham (2013; N=3) Wallis & Guthrie (2024; N not specified) Metcalfe et al. (2020; N=1)	Majority of studies rated 'fair' quality in reviews	Sensitivity and specificity fair. Low predictive validity and moderate discriminant validity	Widely available Brief Multiple translations available Two-stage version with follow-up interview has higher positive predictive value but is more resource-intensive	Performance differs despite range of ages, settings and contexts studied Performance in clinical settings poorer than research settings Caution should be applied in use with children with additional disabilities (for example, hearing/visual impairments and Down syndrome) due to higher rates of false positives and false negatives in these groups
SCQ	Chesnut et al. (2017; N=17) Metcalfe et al. (2020; N=6) Hollocks et al. (2019; N=1) Charman et al. (2007; N=1)	Relatively high amount of research conducted in people with an intellectual disability	Very mixed findings for sensitivity and specificity Low to moderate predictive validity Low inter-rater reliability between parent and teacher scores	Costs associated with licence required to use Ability to capture current and past presentation may be helpful for understanding patient history	Performs better in general population samples than clinical samples Caution should be applied in use with those with behavioural difficulties, intellectual disabilities and hearing/visual impairments, as performance is weaker in these groups Differences in performance identified depending on gender identity, and the country where the assessment is conducted
SRS-2	Summerill et al. (2024; N=6) Baghdadli et al. (2017; N=3) Bruni (2014; N not specified) Thabtah & Peebles (2019; N not specified)	High risk of bias from sampling methods and tests used with often conflicting evidence	Moderate but limited evidence for sensitivity and specificity Low to moderate predictive validity	Wide range of versions adapted for different age groups Experiments that altered the framing of questions found there may be differences in	More reliable for screening in children than adults Evidence of some gender bias

			Satisfactory reliability, but lack of test-retest data	how respondents interpret, depending on context	Caution should be applied in those with an intellectual disability, behavioural difficulties and traits of psychoticism and distress, with the tool having weaker performance for these groups No satisfactory evidence for performance in older adults
ASBQ	Baghdadli et al. (2017; N=1) Horwitz et al. (2016; N=1)	Concerns in reviews about comparison tests used	Satisfactory reliability and validity from very limited evidence Limited data for predictive accuracy	Insufficient evidence identified	Limited research suggests successful differentiation of autism from depression, schizophrenia and ADHD
ASSQ	Kopp & Gillberg (2011; N=1) Thabtah & Peebles (2019; N=2) Påhlman et al. (2020; N=1)	Insufficient evidence identified	Evidence for sensitivity and specificity is high, but limited	Lower completion rates by parents of children with complex disability, suggesting some issues with accessibility or acceptability	Good ability to differentiate autism from ADHD No significant gender differences detected Found to have reasonable sensitivity in those with cerebral palsy
SACS-R	Barbaro et al. (2022; N=1) Mozolic-Staunton et al. (2020; N=1)	Large population sample sizes, may be more suited to population screening than targeted screening	Sensitivity moderate to high Specificity high	Trained clinician needed for SACS-R App version freely available and usable by parents, may reduce demand on services by allowing parents to self-screen their children and aid with decision-making	Insufficient evidence identified

Table 6: Evidence for use of general neurodevelopmental assessment tools for screening and triage in the autism assessment pathway.

General neuro-developmental assessment tools	**Evidence for use in autism screening** (N = number of studies included in paper)	**Research design characteristics**	**Validity and reliability**	**Feasibility and acceptability**	**Consistency across populations/co-occurring conditions/use cases**
A&SQ	Muthusamy et al. (2022; N=43) Hardy et al. (2015; N=1)	Moderate risk of bias from samples and comparison tests used, with mixed designs across studies	Sensitivity and specificity moderate to high Reasonable discriminant validity High correlations with M-CHAT outcomes and diagnostic outcomes	Brief and inexpensive	Higher accuracy for children at older end of age bracket, may be less suitable for younger children in recommended age range
DAWBA	Coscini et al. (2020; N=7 in literature review, 1 original study) McEwen et al. (2016; N=1)	Insufficient evidence identified	Mixed findings for sensitivity and specificity; reducing the cut-off score reduces false negatives but increases false positives Poor inter-rater reliability High sensitivity and specificity for computer algorithm version, ratings correlated with clinician scores. Good ability to predict diagnostic outcome when used with ADOS	Freely available, review by clinician recommended for evaluating (as there are no recommended cut-off scores)	Insufficient evidence identified
CCC	Charman et al. (2007; N=1)	Insufficient evidence identified	Limited evidence for CCC in autism screening suggests high sensitivity and low specificity Good discriminant validity to distinguish autism from ADHD at screening stage	Insufficient evidence identified	No notable difference in performance in those with and without an intellectual disability, suggesting it may be suitable for use in this population, though evidence is very limited

					The tool has lower specificity in those with behavioural problems
SDQ	Warnick et al. (2008; N=3) Russell et al. (2013; N=1) Findon et al. (2016; N=1) Påhlman et al. (2020; N=1) Turcan et al. (2024; N=1)	Even proportions of male/female participants from wide range of backgrounds	Moderate sensitivity and high specificity Low inter-rater reliability for parent and teacher scores, highest specificity in parent-report version compared to self-report and teacher-report	Lower completion rates by parents of children with complex disability, suggesting some issues with accessibility or acceptability	Pilot study found acceptable validity and reliability in use as a screener for co-occurring conditions, but other studies show low ability to discriminate autism from ADHD Consistent performance across settings Poorer performance in autistic adolescents
Early Year's Foundation Stage Profile	Wright et al. (2019; N=1)	Insufficient evidence identified	Scores show correlation with diagnostic outcomes	Carried out routinely by teachers	
Developmental Behaviour Checklist – Autism Screening Algorithm	Brereton et al. (2002; N=1) Witwer & Lacavalier (2009; N=1) Thabtah & Peebles (2019; N not specified)	Risk of bias due to small samples	Good sensitivity but poor specificity	Insufficient evidence identified	Found to be effective in children with an intellectual disability Reduced performance in those with behavioural problems
Conners	Collett et al. (2003; N not specified) Smyth et al. (2016; N=14)	Lack of evidence on use in screening and/or triage in the autism assessment pathway	Insufficient evidence identified	Insufficient evidence identified	Shows mixed sensitivity, specificity, validity and reliability for ADHD identification Evidence of gender bias and overlap with other conditions
BRIEF	Walker & D'Amato (2006; N not specified)	Insufficient evidence identified	Moderate to high reliability	Insufficient evidence identified	Scores are affected by age and gender, but these are adjusted for if scored using recommended methods Research has found little variation in scores dependent on ethnicity

8. Conclusion

There is a wide range of tools available for aiding decisions around screening and triage of referrals in the autism assessment pathway. This report highlights some of the criteria that should be considered when selecting the most appropriate tools for a service to use and evaluates 17 example tools against these criteria. Self-assessment questions are included to support services in making decisions about whether to use tools that are not covered in this report. When choosing between tools, decision-making should remain mindful of contextual factors, including the demographics of patients accessing the service, sources of referrals, the need for transdiagnostic approaches by multidisciplinary teams, and the feasibility and impacts of using tools within the contexts of local NHS pressures.

Services should use tools (or a combination of tools) supported by rigorous research evidence wherever possible, although this report identified substantial gaps in the evidence base. In particular, the use of standardised tools in triage of referrals has rarely been considered, compared to their use in screening. There are also gaps in the evidence related to exploring the applicability of tools across subgroups (for example, based on patient demographics, or clinical need) and the ability of tools to identify alternative or co-occurring conditions. More studies are needed to robustly evaluate tools against the criteria in realistic settings, specifically the context of current NHS autism assessment pathways.

There are many other tools available which were not selected for the current report. Where providers wish to consider these, they should review the available evidence regarding the ERG's recommended criteria to assess their suitability. Service providers should be aware that bespoke tools designed by services to give a broad neurodevelopmental profile are lacking in evidence for their validity, as well as other criteria such as their reliability and consistency across populations and contexts.

Out of the example tools chosen by the ERG for evaluation in this report, there was no single tool that scored highly on all identified criteria. Using tools in combination with each other may help to ameliorate some of the weaknesses of tools, although overreliance on the output of tools should be avoided. Screening and triage decisions in the autism assessment pathway should be supported by clinical expertise.

8.1 Limitations

This report was conducted within a 6-month timeframe from conception to completion and the use of systematic review methods were out of scope for this project. While we aimed to evaluate rigorous research, prioritising previous reviews and meta-analyses, some important findings from individual studies, grey literature and service evaluations may have been missed. Findings may be affected by publication bias. Only studies written in English were included, with UK-based studies prioritised for inclusion. Engagement of an ERG strengthened the understanding of different tools and methods, but the group's availability varied, so unanimous consensus on the selection of tools and criteria was not always possible. Perspectives may also have been missed from disciplines that were not represented within the group. It is essential to examine the wider research evidence (with reference to the appraisal criteria set out here) when evaluating tools for usage in screening and triage that have not been included in this report.

9. References

Allison, C., Auyeung, B., & Baron-Cohen, S. (2012). Toward brief “red flags” for autism screening: The short Autism-Spectrum Quotient and the short quantitative checklist in 1,000 cases and 3,000 controls. *Journal of the American Academy of Child & Adolescent Psychiatry*, *51*(2), 202-212.e7. https://doi.org/10.1016/j.jaac.2011.11.003

Allison, C., Matthews, F. E., Ruta, L., Pasco, G., Soufer, R., Brayne, C., ... Baron-Cohen, S. (2021). Quantitative Checklist for Autism in Toddlers (Q-CHAT). A population screening study with follow-up: The case for multiple time-point screening for autism. *BMJ Paediatrics Open*, *5*(1), e000700. https://doi.org/10.1136/bmjpo-2020-000700

Ashwood, K. L., Gillan, N., Horder, J., Hayward, H., Woodhouse, E., McEwen, F. S., … Murphy, D. G. (2016). Predicting the diagnosis of autism in adults using the Autism-Spectrum Quotient (AQ) questionnaire. *Psychological Medicine*, *46*(12), 2595–2604. https://doi.org/10.1017/S0033291716001082

Astle, D. E., Holmes, J., Kievit, R., & Gathercole, S. E. (2022). Annual Research Review: The transdiagnostic revolution in neurodevelopmental disorders. *Journal of Child Psychology and Psychiatry*, *63*(4), 397–417. https://doi.org/10.1111/jcpp.13481

Apperly, I. A., Lee, R., van der Kleij, S. W., & Devine, R. T. (n.d.). A transdiagnostic approach to neurodiversity in a representative population sample: The N+ 4 model. *JCPP Advances*, e12219. https://doi.org/10.1002/jcv2.12219

Autism Network Scotland (Director). (2020, October 5). *NHS Fife—Fife Neurodevelopmental Questionnaire (F-NDQ)*. https://www.youtube.com/watch?v=Qf0PDoUItOQ

Baghdadli, A., Russet, F., & Mottron, L. (2017). Measurement properties of screening and diagnostic tools for autism spectrum adults of mean normal intelligence: A systematic review. *European Psychiatry*, *44*, 104–124. https://doi.org/10.1016/j.eurpsy.2017.04.009

Barbaro, J., Sadka, N., Gilbert, M., Beattie, E., Li, X., Ridgway, L., ... Dissanayake, C. (2022). Diagnostic accuracy of the Social Attention and Communication Surveillance–Revised with preschool tool for early autism detection in very young children. *JAMA Network Open*, *5*(3), e2146415. https://doi.org/10.1001/jamanetworkopen.2021.46415

Barbaro, J., & Yaari, M. (2020). Study protocol for an evaluation of ASDetect—A Mobile application for the early detection of autism. *BMC Pediatrics*, *20*(1), 21. https://doi.org/10.1186/s12887-019-1888-6

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The Autism-Spectrum Quotient (AQ): Evidence from Asperger Syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders*, *31*(1), 5–17. https://doi.org/10.1023/A:1005653411471

Belcher, H. L., Uglik-Marucha, N., Vitoratou, S., Ford, R. M., & Morein-Zamir, S. (2023). Gender bias in autism screening: Measurement invariance of different model frameworks of the Autism Spectrum Quotient. *BJPsych Open*, *9*(5), e173. https://doi.org/10.1192/bjo.2023.562

Bishop, D. V. M. (1998). Development of the Children's Communication Checklist (CCC): A method for assessing qualitative aspects of communicative impairment in children. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, *39*(6), 879–891. https://doi.org/10.1017/S0021963098002832

Brereton, A. V., Tonge, B. J., Mackinnon, A. J., & Einfeld, S. L. (2002). Screening young people for autism with the Developmental Behavior Checklist. *Journal of the American Academy of Child & Adolescent Psychiatry, 41*(11), 1369–1375. https://doi.org/10.1097/00004583-200211000-00019

Brugha, T., Tyrer, F., Leaver, A., Lewis, S., Seaton, S., Morgan, Z., ... van Rensburg, K. (2020). Testing adults by questionnaire for social and communication disorders, including autism spectrum disorders, in an adult mental health service population. *International Journal of Methods in Psychiatric Research*, *29*(1), e1814. https://doi.org/10.1002/mpr.1814

Bruni, T. P. (2014). Test Review: Social Responsiveness Scale–Second Edition (SRS-2). *Journal of Psychoeducational Assessment*, *32*(4), 365–369. https://doi.org/10.1177/0734282913517525

Campbell, J. M., Klugar, M., Ding, S., Carmody, D. P., Hakonsen, S. J., Jadotte, Y. T., ... Munn, Z. (2015). Diagnostic test accuracy: Methods for systematic review and meta-analysis. J*BI Evidence Implementation, 13*(3), 154. https://doi.org/10.1097/XEB.0000000000000061

Charman, T., Baird, G., Simonoff, E., Loucas, T., Chandler, S., Meldrum, D., & Pickles, A. (2007). Efficacy of three screening instruments in the identification of autistic-spectrum disorders. *The British Journal of Psychiatry*, *191*(6), 554–559. https://doi.org/10.1192/bjp.bp.107.040196

Charman, T., & Gotham, K. (2013). Measurement Issues: Screening and diagnostic instruments for autism spectrum disorders – lessons from research and practise. *Child and Adolescent Mental Health*, *18*(1), 52–63. https://doi.org/10.1111/j.1475-3588.2012.00664.x

Chesnut, S. R., Wei, T., Barnard-Brak, L., & Richman, D. M. (2017). A meta-analysis of the Social Communication Questionnaire: Screening for autism spectrum disorder. *Autism*, *21*(8), 920–928. https://doi.org/10.1177/1362361316660065

Collett, B. R., Ohan, J. L., & Myers, K. M. (2003). Ten-year review of rating scales. V: Scales assessing attention-deficit/hyperactivity disorder. *Journal of the American Academy of Child & Adolescent Psychiatry*, *42*(9), 1015–1037. https://doi.org/10.1097/01.CHI.0000070245.24125.B6

Conner, C. M., Cramer, R. D., & McGonigle, J. J. (2019). Examining the diagnostic validity of autism measures among adults in an outpatient clinic sample. *Autism in Adulthood*, *1*(1), 60–68. https://doi.org/10.1089/aut.2018.0023

Constantino, J. N. (2013). Social Responsiveness Scale. In F. R. Volkmar (Ed.), *Encyclopedia of Autism Spectrum Disorders* (pp. 2919–2929). Springer. https://doi.org/10.1007/978-1-4419-1698-3_296

Coscini, N., Srinivasan, R., & Skuse, D. (2020). Validating the Developmental and Well-Being Assessment (DAWBA) in a clinical population with high-functioning autism. *F1000Research*. [version 1; peer review: 2 approved with reservations], *9*(622). https://doi.org/10.12688/f1000research.24409.1

Derks, O., Heinrich, M., Brooks, W., Sterkenburg, P., McCarthy, J., Underwood, L., & Sappok, T. (2017). The Social Communication Questionnaire for adults with intellectual disability: SCQ-AID. *Autism Research*, *10*(9), 1481–1490. https://doi.org/10.1002/aur.1795

Ehlers, S., Gillberg, C., & Wing, L. (1999). A screening questionnaire for Asperger Syndrome and other high-functioning autism spectrum disorders in school age children. *Journal of Autism and Developmental Disorders*, *29*(2), 129–141. https://doi.org/10.1023/A:1023040610384

Findon, J., Cadman, T., Stewart, C. S., Woodhouse, E., Eklund, H., Hayward, H., ... McEwen, F. S. (2016). Screening for co-occurring conditions in adults with autism spectrum disorder using the strengths and difficulties questionnaire: A pilot study. *Autism Research*, *9*(12), 1353–1363. https://doi.org/10.1002/aur.1625

Genovesi, E., Ullmer, P., Bhatti, L., Meyer, P., Memon, P., Panchani, D., ... Hoekstra, R. A. (2023). Perspectives on children's autistic traits in UK-based British and Egyptian/Sudanese communities. *Research in Developmental Disabilities*, *140*, 104576. https://doi.org/10.1016/j.ridd.2023.104576

Gernsbacher, M. A., Stevenson, J. L., & Dern, S. (2017). Specificity, contexts, and reference groups matter when assessing autistic traits. *PLOS ONE*, *12*(2), e0171931. https://doi.org/10.1371/journal.pone.0171931

Gioia, G. A., Isquith, P. K., Guy, S. C., & Kenworthy, L. (2000). *Behavior Rating Inventory of Executive Function®, Second Edition (BRIEF® 2)*. https://www.nzcer.org.nz/system/files/BRIEF2%20Fact%20Sheet.pdf

Goodman, R. (1999). The extended version of the strengths and difficulties questionnaire as a guide to child psychiatric caseness and consequent burden. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, *40*(5), 791–799. https://doi.org/10.1111/1469-7610.00494

Gray, K. M., Tonge, B. J., Sweeney, D. J., & Einfeld, S. L. (2008). Screening for autism in young children with developmental delay: An evaluation of the Developmental Behaviour Checklist: Early Screen. *Journal of Autism and Developmental Disorders*, *38*(6), 1003–1010. https://doi.org/10.1007/s10803-007-0473-2

Green, J., Leadbitter, K., Ainsworth, J., & Bucci, S. (2022). An integrated early care pathway for autism. *The Lancet Child & Adolescent Health*, *6*(5), 335–344. https://doi.org/10.1016/S2352-4642(22)00037-2

Hardy, S., Haisley, L., Manning, C., & Fein, D. (2015). Can screening with the Ages and Stages Questionnaire detect autism? *Journal of Developmental and Behavioral Pediatrics*, *36*(7), 536–543. https://doi.org/10.1097/DBP.0000000000000201

Hollocks, M. J., Casson, R., White, C., Dobson, J., Beazley, P., & Humphrey, A. (2019). Brief report: An evaluation of the Social Communication Questionnaire as a screening tool for autism spectrum disorder in young people referred to child & adolescent mental health services. *Journal of Autism and Developmental Disorders*, *49*(6), 2618–2623. https://doi.org/10.1007/s10803-019-03982-6

Horwitz, E. H., Schoevers, R. A., Ketelaars, C. E. J., Kan, C. C., van Lammeren, A. M. D. N., Meesters, Y., ... Hartman, C. A. (2016). Clinical assessment of ASD in adults using self- and other-report: Psychometric properties and validity of the Adult Social Behavior Questionnaire (ASBQ). *Research in Autism Spectrum Disorders*, *24*, 17–28. https://doi.org/10.1016/j.rasd.2016.01.003

Hudson, C. C., Bajwa, H. E., Beard, C., & Björgvinsson, T. (2024). Psychometric properties of the 10-item Autism Spectrum Quotient in an acute psychiatric sample. *Research in Autism Spectrum Disorders*, *110*, 102299. https://doi.org/10.1016/j.rasd.2023.102299

Jones, S. L., Johnson, M., Alty, B., & Adamou, M. (2021). The effectiveness of RAADS-R as a screening tool for adult ASD populations. *Autism Research and Treatment*, *2021*, e9974791. https://doi.org/10.1155/2021/9974791

Kopp, S., & Gillberg, C. (2011). The Autism Spectrum Screening Questionnaire (ASSQ)-Revised Extended Version (ASSQ-REV): An instrument for better capturing the autism phenotype in girls? A preliminary study involving 191 clinical cases and community controls. *Research in Developmental Disabilities*, *32*(6), 2875–2888. https://doi.org/10.1016/j.ridd.2011.05.017

Kraijer, D., & de Bildt, A. (2005). The PDD-MRS: An instrument for identification of autism spectrum disorders in persons with mental retardation. *Journal of Autism and Developmental Disorders, 35*(4), 499–513. https://doi.org/10.1007/s10803-005-5040-0

Lai, M. C., Kassee, C., Besney, R., Bonato, S., Hull, L., Mandy, W., Szatmari, P., & Ameis, S. H. (2019). Prevalence of co-occurring mental health diagnoses in the autism population: a systematic review and meta-analysis. *The Lancet Psychiatry*, *6*(10), 819–829. https://doi.org/10.1016/S2215-0366(19)30289-5

Leekam, S. R., Libby, S. J., Wing, L., Gould, J., & Taylor, C. (2002). The Diagnostic Interview for Social and Communication Disorders: Algorithms for ICD-10 childhood autism and Wing and

Gould autistic spectrum disorder. *Journal of Child Psychology and Psychiatry*, *43*(3), 327–342. https://doi.org/10.1111/1469-7610.00024

Levy, S. E., Wolfe, A., Coury, D., Duby, J., Farmer, J., Schor, E., ... Warren, Z. (2020). Screening tools for autism spectrum disorder in primary care: A systematic evidence review. *Pediatrics*, *145*(Supplement_1), S47–S59. https://doi.org/10.1542/peds.2019-1895H

Livanis, A., & Mouzakitis, A. (2010). The treatment validity of autism screening instruments. *Assessment for Effective Intervention*, *35*(4), 206–217. https://doi.org/10.1177/1534508410381041

McEwen, F. S., Stewart, C. S., Colvert, E., Woodhouse, E., Curran, S., Gillan, N., ... Bolton, P. (2016). Diagnosing autism spectrum disorder in community settings using the Development and Well-Being Assessment: Validation in a UK population-based twin sample. *Journal of Child Psychology and Psychiatry*, *57*(2), 161–170. https://doi.org/10.1111/jcpp.12447

McQuaid, G. A., Ratto, A. B., Jack, A., Khuu, A., Smith, J. V., Duane, S. C., ... Strang, J. F. (2024). Gender, assigned sex at birth, and gender diversity: Windows into diagnostic timing disparities in autism. *Autism*, 13623613241243117. https://doi.org/10.1177/13623613241243117

Metcalfe, D., McKenzie, K., McCarty, K., & Murray, G. (2020). Screening tools for autism spectrum disorder, used with people with an intellectual disability: A systematic review. *Research in Autism Spectrum Disorders*, *74*, 101549. https://doi.org/10.1016/j.rasd.2020.101549

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., ... de Vet, H. C. W. (2010). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international Delphi study. *Quality of Life Research*, *19*(4), 539–549. https://doi.org/10.1007/s11136-010-9606-8

Mozolic-Staunton, B., Donelly, M., Yoxall, J., & Barbaro, J. (2020). Early detection for better outcomes: Universal developmental surveillance for autism across health and early childhood education settings. *Research in Autism Spectrum Disorders*, *71*, 101496. https://doi.org/10.1016/j.rasd.2019.101496

Muggleton, J. (2024) The Fife Neurodevelopmental Questionnaire Child Version 2. Available at www.moodcafe.co.uk/f-ndq/

Murray, A. L., Allison, C., Smith, P. L., Baron-Cohen, S., Booth, T., & Auyeung, B. (2017). Investigating diagnostic bias in autism spectrum conditions: An item response theory analysis of sex bias in the AQ-10. *Autism Research*, *10*(5), 790–800. https://doi.org/10.1002/aur.1724

Muthusamy, S., Wagh, D., Tan, J., Bulsara, M., & Rao, S. (2022). Utility of the Ages and Stages Questionnaire to identify developmental delay in children aged 12 to 60 months: A systematic review and meta-analysis. *JAMA Pediatrics*, *176*(10), 980–989. https://doi.org/10.1001/jamapediatrics.2022.3079

National Institute for Health and Care Excellence (NICE). (2017, December 20). *Recommendations | Autism spectrum disorder in under 19s: Recognition, referral and diagnosis | Guidance | NICE*. NICE. https://www.nice.org.uk/guidance/cg128/chapter/Recommendations#recognising-children-and-young-people-who-may-be-autistic

National Institute for Health and Care Excellence (NICE). (2021, June 14). *Recommendations | Autism spectrum disorder in adults: Diagnosis and management | Guidance | NICE*. NICE. https://www.nice.org.uk/guidance/CG142/chapter/Recommendations#identification-and-assessment-2

NHS Digital. (2023, December 13). *Autism Statistics, October 2022 to September 2023*. NHS Digital. https://digital.nhs.uk/data-and-information/publications/statistical/autism-statistics/october-2022-to-september-2023

NHS England. (2023a). *A national framework to deliver improved outcomes in all-age autism assessment pathways: Guidance for integrated care boards*. https://www.england.nhs.uk/long-read/a-national-framework-to-deliver-improved-outcomes-in-all-age-autism-assessment-pathways-guidance-for-integrated-care-boards/

NHS England. (2023b). *Operational guidance to deliver improved outcomes in all-age autism assessment pathways: Guidance for integrated care boards*. https://www.england.nhs.uk/long-read/operational-guidance-to-deliver-improved-outcomes-in-all-age-autism-assessment-pathways-guidance-for-integrated-care-boards/#stage-2-screening-and-triage

Nishiyama, T., Suzuki, M., Adachi, K., Sumi, S., Okada, K., Kishino, H., ... Kanne, S. M. (2014). Comprehensive comparison of self-administered questionnaires for measuring quantitative autistic traits in adults. *Journal of Autism and Developmental Disorders*, *44*(5), 993–1007. https://doi.org/10.1007/s10803-013-2020-7

O'Nions, E., Petersen, I., Buckman, J. E. J., Charlton, R., Cooper, C., Corbett, A., ... Stott, J. (2023). Autism in England: Assessing underdiagnosis in a population-based cohort study of prospectively collected primary care data. *The Lancet Regional Health – Europe*, *29*. https://doi.org/10.1016/j.lanepe.2023.100626

Påhlman, M., Gillberg, C., Wentz, E., & Himmelmann, K. (2020). Autism spectrum disorder and attention-deficit/hyperactivity disorder in children with cerebral palsy: Results from screening in a population-based group. *European Child & Adolescent Psychiatry*, *29*(11), 1569–1579. https://doi.org/10.1007/s00787-020-01471-1

Perera, H., Jeewandara, K. C., Seneviratne, S., & Guruge, C. (2017). Culturally adapted pictorial screening tool for autism spectrum disorder: A new approach. *World Journal of Clinical Pediatrics*, *6*(1), 45–51. https://doi.org/10.5409/wjcp.v6.i1.45

Ritvo, R. A., Ritvo, E. R., Guthrie, D., Ritvo, M. J., Hufnagel, D. H., McMahon, W., ... Eloff, J. (2011). The Ritvo Autism Asperger Diagnostic Scale-Revised (RAADS-R): A scale to assist the

diagnosis of autism spectrum disorder in adults: An international validation study. *Journal of Autism and Developmental Disorders*, *41*(8), 1076–1089. https://doi.org/10.1007/s10803-010-1133-5

Robins, D. L., Fein, D., Barton, M. L., & Green, J. A. (2001). The Modified Checklist for Autism in Toddlers: An initial study investigating the early detection of autism and pervasive developmental disorders. *Journal of Autism and Developmental Disorders*, *31*(2), 131–144. https://doi.org/10.1023/A:1010738829569

Russell, G., Rodgers, L. R., & Ford, T. (2013). The Strengths and Difficulties Questionnaire as a predictor of parent-reported diagnosis of autism spectrum disorder and attention deficit hyperactivity disorder. *PLOS ONE*, *8*(12), e80247. https://doi.org/10.1371/journal.pone.0080247

Russell, G., Stapley, S., Newlove-Delgado, T., Salmon, A., White, R., Warren, F., ... Ford, T. (2022). Time trends in autism diagnosis over 20 years: A UK population-based cohort study. *Journal of Child Psychology and Psychiatry*, *63*(6), 674–682. https://doi.org/10.1111/jcpp.13505

Ruzich, E., Allison, C., Smith, P., Watson, P., Auyeung, B., Ring, H., & Baron-Cohen, S. (2015). Measuring autistic traits in the general population: a systematic review of the Autism-Spectrum Quotient (AQ) in a nonclinical population sample of 6,900 typical adult males and females. *Molecular autism*, *6*, 2. https://doi.org/10.1186/2040-2392-6-2

Santos, C. L. dos, Barreto, I. I., Floriano, I., Tristão, L. S., Silvinato, A., & Bernardo, W. M. (2024). Screening and diagnostic tools for autism spectrum disorder: Systematic review and meta-analysis. *Clinics*, *79*, 100323. https://doi.org/10.1016/j.clinsp.2023.100323

Sappok, T., Gaul, I., Bergmann, T., Dziobek, I., Bölte, S., Diefenbacher, A., & Heinrich, M. (2014). The Diagnostic Behavioral Assessment for autism spectrum disorder—Revised: A screening instrument for adults with intellectual disability suspected of autism spectrum disorders. *Research in Autism Spectrum Disorders*, *8*(4), 362–375. https://doi.org/10.1016/j.rasd.2013.12.016

Smyth, A. C., & Meier, S. T. (2019). Evaluating the psychometric properties of the Conners Adult ADHD Rating Scales. *Journal of Attention Disorders*, *23*(10), 1111–1118. https://doi.org/10.1177/1087054715624230

Snow, A. (2013). Social Communication Questionnaire. In F. R. Volkmar (Ed.), *Encyclopedia of Autism Spectrum Disorders* (pp. 2893–2895). Springer. https://doi.org/10.1007/978-1-4419-1698-3_1651

Stacey, R., & Cage, E. (2023). “Simultaneously vague and oddly specific”: Understanding autistic people’s experiences of decision making and research questionnaires. *Autism in Adulthood*, *5*(3), 263–274. https://doi.org/10.1089/aut.2022.0039

Sturm, A., Huang, S., Bal, V., & Schwartzman, B. (2024). Psychometric exploration of the RAADS-R with autistic adults: Implications for research and clinical practice. *Autism*, 13623613241228329. https://doi.org/10.1177/13623613241228329

Summerill, J., Leedale, C., Thompson, D., Ingham, B., & Rutter, L. (2024). Autism assessment and diagnosis in older adults: A systematic review of measures and tools. Author's accepted manuscript sent via private email correspondence.

Thabtah, F., & Peebles, D. (2019). Early autism screening: A comprehensive review. *International Journal of Environmental Research and Public Health*, *16*(18), Article 18. https://doi.org/10.3390/ijerph16183502

Tollerfield, I., & Pearce, H. (2020). Use of the Thinking Patterns in Autism Profiling Model within a diagnostic assessment service for autism. *Good Autism Practice (GAP)*, *21*(2), 99–117. https://www.ingentaconnect.com/contentone/bild/gap/2020/00000021/00000002/art00009

Tromans, S., Chester, V., Gemegah, E., Roberts, K., Morgan, Z., Yao, G. L., & Brugha, T. (2020). Autism identification across ethnic groups: A narrative review. *Advances in Autism*, *7*(3), 241–255. https://doi.org/10.1108/AIA-03-2020-0017

Turcan, C., Delamain, H., Loke, A., Pender, R., Mandy, W., & Saunders, R. (2024). Measurement invariance of the parent-reported Strengths and Difficulties Questionnaire in autistic adolescents. *Autism*, 13623613241236805. https://doi.org/10.1177/13623613241236805

UK National Screening Committee. (2022). *Autism—UK National Screening Committee (UK NSC)—GOV.UK*. GOV.UK. https://view-health-screening-recommendations.service.gov.uk/autism/

Waldren, L. H., Clutterbuck, R. A., & Shah, P. (2021). Erroneous NICE guidance on autism screening. *The Lancet Psychiatry*, *8*(4), 276–277. https://doi.org/10.1016/S2215-0366(21)00065-1

Walker, J. M., & D'Amato, R. C. (2006). Test review: Behavior Rating Inventory of Executive Function-Self-Report version. *Journal of Psychoeducational Assessment*, *24*(4), 394–398. https://doi.org/10.1177/0734282906288390

Wallis, K. E., & Guthrie, W. (2024). Screening for autism: A review of the current state, ongoing challenges, and novel approaches on the horizon. *Pediatric Clinics of North America, 71*(2), 127–155. https://doi.org/10.1016/j.pcl.2023.12.003

Warnick, E. M., Bracken, M. B., & Kasl, S. (2008). Screening efficiency of the Child Behavior Checklist and Strengths and Difficulties Questionnaire: A systematic review. *Child and Adolescent Mental Health*, *13*(3), 140–147. https://doi.org/10.1111/j.1475-3588.2007.00461.x

Wigham, S., Rodgers, J., Berney, T., Le Couteur, A., Ingham, B., & Parr, J. R. (2019). Psychometric properties of questionnaires and diagnostic measures for autism spectrum disorders in adults: A systematic review. *Autism*, *23*(2), 287–305. https://doi.org/10.1177/1362361317748245

Witwer, A. N., & Lecavalier, L. (2007). Autism screening tools: An evaluation of the Social Communication Questionnaire and the Developmental Behaviour Checklist–Autism Screening Algorithm. *Journal of Intellectual & Developmental Disability*, *32*(3), 179–187. https://doi.org/10.1080/13668250701604776

Wright, B., Mon-Williams, M., Kelly, B., Williams, S., Sims, D., Mushtaq, F., ... Wright, J. (2019). Investigating the association between early years foundation stage profile scores and subsequent diagnosis of an autism spectrum disorder: A retrospective study of linked healthcare and education data. *BMJ Paediatrics Open*, *3*(1), e000483. https://doi.org/10.1136/bmjpo-2019-000483

10. Appendices

Appendix 1. How we produced this document

Autistica

This report was authored by Autistica. Autistica is the UK's leading autism research and campaigning charity. Our mission is to create breakthroughs that enable all autistic people to live happier, healthier, longer lives. We do this by funding research, shaping policy and working with autistic people to make more of a difference.

With thanks to the Autistica Insight Group, and ERG experts:

Sandy Bering, Strategic Lead Clinical Commissioner, Greater Manchester Health & Social Care Partnership

Professor Terry Brugha, Professor of Psychiatry, University of Leicester

Dr Anna Crabtree, Principal Clinical Psychologist, Royal Holloway University of London

Dr Will Farr, Assistant Professor of Education, University of Cambridge

Fay Hendry, Occupational Therapist, South London and Maudsley NHS Foundation Trust

Dr Rosa Hoekstra, Reader in Global Perspectives on Neurodevelopmental Disabilities, King's College London

Professor Patricia Howlin, Emeritus Professor, King's College London

Sarah Jackson, Consultant Nurse, Lancashire & South Cumbria NHS Foundation Trust

Professor Will Mandy, Professor of Neurodevelopmental Conditions, University College London

Professor Mark Mon-Williams, Professor of Psychology, University of Leeds

Dr Raja Mukherjee, Consultant Psychiatrist, Surrey & Borders NHS Foundation Trust

Dr Heidi Phillips, Neurodevelopmental Specialist GP, University of Wales

Professor Andrew Pickles, Professor of Biostatistics and Psychological Methods, King's College London

Dr Venkat Reddy, Consultant Neurodevelopmental Paediatrician, Cambridgeshire & Peterborough NHS Foundation Trust

Dr Marion Rutherford, National Autism Implementation Team Research Lead, Queen Margaret University

Dr Louise Rutter, Principal Clinical Psychologist, Cumberland, Northumbria, Tyne & Wear NHS Foundation Trust

Dr Sam Tromans, Associate Professor of Psychiatry, University of Leicester

Dr Lucy Waldren, Researcher and Teaching Associate in Psychology, University of Bath

Ann Wilson, Adult Autism Service Clinical Lead, Berkshire Healthcare NHS Foundation Trust

Appendix 2. List of Abbreviations

A&SQ	Ages and Stages Questionnaire
ADHD	Attention Deficit Hyperactivity Disorder
ADOS	Autism Diagnostic Observation Schedule
AQ	Autism Spectrum Quotient
ASBQ	Adult Social Behaviour Questionnaire
ASSQ	Autism Spectrum Screening Questionnaire
AUC	Area Under Curve
BRIEF	Behavior Rating Inventory of Executive Function
CCC	Children's Communication Checklist
CCG	Clinical Commissioning Group
DAWBA	Developmental and Wellbeing Assessment
DBC	Developmental Behaviour Checklist
ERG	Expert Reference Group
ICB	Integrated Care Board
JBI	Joanna Briggs Institute
M-CHAT	Modified Checklist for Autism in Toddlers
NICE	National Institute for Health and Care Excellence
NPV	Negative Predictive Value
NSC	National Screening Committee
PPV	Positive Predictive Value
RAADS-R	Ritvo Asperger and Autism Diagnostic Scale-Revised
SACS-R	Social Attention Communication Surveillance-Revised
SCQ	Social Communication Questionnaire
SDQ	Strengths and Difficulties Questionnaire
SRS-2	Social Responsiveness Scale 2nd edition

Appendix 3. Glossary

Acceptability	Acceptability concerns how well a tool is received by patients, their families/carers, assessment teams and referrers.
Area Under Curve (AUC)	AUC is a measure of accuracy.
Ceiling effects	Ceiling effects are when everyone scores very high on the measure, which may indicate issues with validity.
Construct validity	Construct validity refers to whether the tool is measuring what it intends to measure.
Correlation	A correlation is a relationship between two variables, whereby a change in one variable is associated with a change in the other. For example, as scores on one measure of autistic traits increase, scores on another measure also increase.
Cross-cultural validity	Cross-cultural validity refers to whether the tool measures what it intends to measure across different cultures.
Discriminant validity	Discriminant validity is the ability of a tool to discriminate between autism and other conditions.
Efficiency	Efficiency refers to whether a tool is effective at its intended aims.
Exclusion criteria	Exclusion criteria are a type of eligibility criteria (alongside inclusion criteria) that specify who *cannot* participate in a study. These comprise characteristics used to identify potential research participants who should *not* be included in a study. For example, some studies may not include autistic people who have a co-occurring intellectual disability.
False negative	In the context of screening for autism, a false negative result is when the output of a tool suggests that the person *is not* likely to be autistic, but an assessment confirms that the person *is* autistic.
False positive	In the context of screening for autism, a false positive result is when the output of a tool suggests that the person *is* likely to be autistic, but an assessment confirms that the person *is not* autistic.
Feasibility	Feasibility refers to whether it is practical and viable to use for a specific purpose.
Floor effects	Floor effects are when everyone scores very low on the measure, which may indicate issues with validity.
Inclusion criteria	Inclusion criteria are a type of eligibility criteria (alongside exclusion criteria) that specify who *can* participate in a study. These comprise characteristics that potential research participants *must* have to be included in a study. For example, a study may specify that autistic children and young people aged 5-18 are included in a study.
Informant-report	Informant-report refers to when a tool is completed by an informant (e.g., a clinician, family member, or education professional) on behalf of the person.
Internal consistency	Internal consistency refers to whether all items in the tool are measuring the same construct.
Inter-rater reliability	Inter-rater reliability refers to when the ratings of different respondents to result in similar scores. For example, when the tool is completed by the patient (self-report), a family member, or different clinicians (informant-report).

Pooled outcomes	Pooled outcomes are the results of multiple studies combined, for example, in a meta-analysis.
Population screening	Population screening involves inviting everyone from a specified population to undertake a brief test to identify a subset who would benefit from more resource intensive or more invasive tests, or from receiving information or intervention.
Predictive validity	Predictive validity is the ability of a tool to predict a diagnostic outcome.
Reliability	Reliability refers to whether the output is consistent when repeated.
Risk of bias	Risk of bias refers to the introduction of bias into studies because of methodological insufficiencies or errors.
Screening	Screening is defined in this report as decisions around whether to refer a person for an assessment or whether the assessment service should accept a referral.
Self-report	Self-report refers to when a person fills out a questionnaire or tool for themselves (instead of informant report, where the tool is filled out by someone else).
Sensitivity	Sensitivity is the ability of a tool to correctly identify people with a condition.
Structural validity	Structural validity is defined as the degree to which scores on a measure are an accurate reflection of the construct being measured (e.g., the extent to which a tool designed to assess autistic traits is an accurate reflection of the diagnostic criteria for autism).
Targeted screening	Targeted screening is when people are screened for a condition because they have been identified as being at an increased likelihood of having the condition compared with the general population.
Transdiagnostic	A transdiagnostic tool is a tool which can be applied across multiple conditions, for example, a tool that assesses functioning across multiple neurodevelopmental conditions.
Triage	Triage is defined in the report as decisions around which patients to prioritise for an assessment and identifying which components will be included in an autism assessment.
True negative	In the context of screening for autism, a true negative result is when the output of a tool suggests that the person *is not* likely to be autistic, and an assessment confirms that the person *is not* autistic.
True positive	In the context of screening for autism, a true positive result is when the output of a tool suggests that the person *is* likely to be autistic, and an assessment confirms that the person *is* autistic.
T-score	A t-score is a standardised score on a measure, standardised across the population such that scores sit around a normal distribution. These can be used to determine how a person scores on a measure compared to the population. Some measures may use different standardised scoring for different age groups (for example, age 6-10 and 10-14), to account for differences in performance across age groups.
Validity	Validity refers to whether the tool measures what it intends to measure.

Appendix 4. UK NSC criteria for screening programmes

Population Screening

The below text is copied verbatim from Criteria for a population screening programme

1. The condition should be an important health problem as judged by its frequency and/or severity. The epidemiology, incidence, prevalence and natural history of the condition should be understood, including development from latent to declared disease and/or there should be robust evidence about the association between the risk or disease marker and serious or treatable disease.
2. All the cost-effective primary prevention interventions should have been implemented as far as practicable.
3. If the carriers of a mutation are identified as a result of screening the natural history of people with this status should be understood, including the psychological implications.
4. There should be a simple, safe, precise and validated screening test.
5. The distribution of test values in the target population should be known and a suitable cut-off level defined and agreed.
6. The test, from sample collection to delivery of results, should be acceptable to the target population.
7. There should be an agreed policy on the further diagnostic investigation of individuals with a positive test result and on the choices available to those individuals.
8. If the test is for a particular mutation or set of genetic variants the method for their selection and the means through which these will be kept under review in the programme should be clearly set out.
9. There should be an effective intervention for patients identified through screening, with evidence that intervention at a pre-symptomatic phase leads to better outcomes for the screened individual compared with usual care. Evidence relating to wider benefits of screening, for example those relating to family members, should be taken into account where available. However, where there is no prospect of benefit for the individual screened then the screening programme should not be further considered.
10. There should be agreed evidence based policies covering which individuals should be offered interventions and the appropriate intervention to be offered.
11. There should be evidence from high quality randomised controlled trials that the screening programme is effective in reducing mortality or morbidity. Where screening is aimed solely at providing information to allow the person being screened to make an 'informed choice' (such as Down's syndrome or cystic fibrosis carrier screening), there must be evidence from high quality trials that the test accurately measures risk. The information that is provided about the test and its outcome must be of value and readily understood by the individual being screened.
12. There should be evidence that the complete screening programme (test, diagnostic procedures, treatment/ intervention) is clinically, socially and ethically acceptable to health professionals and the public.
13. The benefit gained by individuals from the screening programme should outweigh any harms, for example from overdiagnosis, overtreatment, false positives, false reassurance, uncertain findings and complications.
14. The opportunity cost of the screening programme (including testing, diagnosis and treatment, administration, training and quality assurance) should be economically balanced in relation to expenditure on medical care as a whole (value for money).

Assessment against this criteria should have regard to evidence from cost benefit and/or cost effectiveness analyses and have regard to the effective use of available resource.

15. Clinical management of the condition and patient outcomes should be optimised in all health care providers prior to participation in a screening programme.
16. All other options for managing the condition should have been considered (such as improving treatment or providing other services), to ensure that no more cost effective intervention could be introduced or current interventions increased within the resources available.
17. There should be a plan for managing and monitoring the screening programme and an agreed set of quality assurance standards.
18. Adequate staffing and facilities for testing, diagnosis, treatment and programme management should be available prior to the commencement of the screening programme.
19. Evidence-based information, explaining the purpose and potential consequences of screening, investigation and preventative intervention or treatment, should be made available to potential participants to assist them in making an informed choice.
20. Public pressure for widening the eligibility criteria for reducing the screening interval, and for increasing the sensitivity of the testing process, should be anticipated. Decisions about these parameters should be scientifically justifiable to the public.

Targeted screening

The below is copied verbatim from criteria for a targeted screening programme

1. The health impact of the condition and its course should be understood, with evidence that serious disease can be identified or predicted by an agreed level of identifiable risk or marker.
2. There should be evidence from appropriately designed studies and models on cost effectiveness for the:
    a) screening test – this should be a simple test that has evidence of suitable accuracy and technical performance derived from studies in the population in which the test is being used,
    b) intervention – there should be better outcomes from early intervention/those at a pre-symptomatic stage, for the screened individual, compared with usual care.
3. There should be a diagnostic investigation available for individuals with a positive test screening result, with evidence that subsequent tests can distinguish those who would benefit from interventions from those who would not.
4. The overall benefits from the screening programme should outweigh the harms, for example, from overdiagnosis, overtreatment, false positives, false reassurance and uncertain findings.
5. There should be a robust and inclusive evidence-based selection criteria for identifying those eligible for targeted screening.
6. There should be evidence that the screening programme is acceptable to the public and health professionals, with appropriately balanced information available to those invited to attend screening.
7. The UK NSC should carry out a feasibility assessment (informed by NHS practice) that indicates the screening programme would be achievable; with evidence for monitoring

and quality assuring the programme, adequate staffing and facilities being available to meet the requirements of programme delivery.

Appendix 5. Literature review

A.3.1 Description of autism-specific tools

The following tools ranked highest in ERG polls for inclusion in the present report.

The **Autism Spectrum Quotient (AQ-50)**, developed by Baron-Cohen et al. (2001), is a descriptive self-report questionnaire for adults covering "social skill, attention switching, attention to detail, communication, imagination" (Baghdadli et al., 2017). Items are scored on a 4-point scale from "definitely disagree" to "definitely agree", with reverse scoring used for half of the questions. Studies have recommended cut-off scores of either 26 or 32 to indicate that an autism assessment would be beneficial, although its creators caution that it is unsuitable for population screening (Ruzich et al., 2015). The **AQ-10** is a 10-item version with a cut-off score of ≥6 (Allison et al., 2012).

The **Ritvo Asperger and Autism Diagnostic Scale-Revised (RAADS-R)**, developed by Ritvo et al. (2011), is an adult self-report questionnaire covering "social relatedness; circumscribed interests; language; sensorimotor and stereotypes" (Baghdadli et al., 2017). Respondents are asked to rate whether traits are present now, when they were younger, both, or not present at all. Some questions use reverse scoring.

The **Modified Checklist for Autism in Toddlers (M-CHAT)** was developed by Robins et al. (2001) and designed to improve the sensitivity of the original CHAT tool. It is a freely available parent-report questionnaire for children of 16-30 months. A version with a follow-up interview is also available (M-CHAT-R/F).

The **Social Communication Questionnaire (SCQ)** by Rutter et al. (2003, in Snow, 2013), which is suitable for children or adults, comprises 40 items on current behaviour and behaviour at age 4-5, with items scored 0 or 1 and a cut-off of ≥15. It is intended to be completed by a parent.

The **Social Responsiveness Scale 2nd edition-Adult Form (SRS-2)** developed by Constantino and Gruber (2012, in Constantino, 2013) is suitable for children and adults, and can be self- or informant-rated. It covers domains of "social awareness, social cognition, social communication, social motivation, restricted interests and repetitive behavior" (Baghdadli et al., 2017), with scores on subscales used to generate an overall T-score and an optimal cut-off score of 70 (Thabtah & Peebles, 2019).

The **Social Attention Communication Surveillance-Revised (SACS-R)** is a tool designed for universal surveillance of children for early detection of autism, such as within health visitor appointments (Mozolic-Staunton et al., 2020). A free app version of the SACS-R tool called ASDetect has been developed for use by parents to indicate their child's likelihood of being autistic, and its psychometric properties are being tested (Barbaro & Yaari, 2020).

The **Adult Social Behaviour Questionnaire (ASBQ)** can be self-completed or completed by an informant, comprising 44 items scored on a three-point scale (Horwitz et al., 2016). The **Autism Spectrum Screening Questionnaire (ASSQ)** is a 27-item informant checklist designed for screening of children and adolescents (Ehlers et al., 1999). The **Past and Present Behaviour Schedule** (Leekham et al., 2008) is an interview or questionnaire for informants, suitable for obtaining a history of adults and children referred for autism assessment, with the outcome based on clinical judgement.

A.3.2 Extent of evidence and design characteristics of studies

A systematic review by Baghdadli et al. (2017) identified 22 studies applying the AQ to adult samples without intellectual disability, 3 studies involving the RAADS-R and 3 involving the SRS-2. A review of the SRS-2 suggested there has been a lack of independent research on this measure (Bruni, 2014). A meta-analysis by Chesnut et al. (2016) identified 17 studies involving the SCQ. Three studies involving the M-CHAT were included in a review by Charman and Gotham (2013). Evidence for the ASBQ is limited (Baghdadli et al., 2017).

The AQ-50 has been evaluated more than the AQ-10. More studies of the AQ have included a higher proportion of males than females, especially for clinical samples. A risk of bias has been identified across studies of the AQ due to small sample sizes and the applicability of index tests used for comparisons. Studies of the RAADS-R have had more even gender ratios but were also at risk of bias from patient selection and index tests used; these were also concerns for these issues among studies of the SRS-2. The ASBQ showed risks of bias from index and reference tests (Baghdadli et al., 2017). Brugha et al. (2020) also criticise studies of the AQ and RAADS-R for using case-control designs, which may lead to overestimation of sensitivity and specificity. No published evidence could be identified for the measurement properties of the Past and Present Behaviour Schedule. The use of systematic search and review methods were out of scope for the current review, and therefore publication bias cannot be fully appraised.

A.3.3 Validity and reliability

Sensitivity and specificity are valued between 0 and 1. A sensitivity value of 1 indicates that 100% of people identified as autistic are true positives, and a specificity value of 1 indicates that 100% of people identified as not autistic are true negatives. Literature suggests the AQ has satisfactory sensitivity but may lack specificity. Estimates for sensitivity range from 0.75-0.95 for the AQ-50 and 0.62-0.88 for the AQ-10, while specificity for the AQ-50 ranges from 0.52-0.97 and for the AQ-10 from 0.66-0.91 (Baghdadli et al., 2017). Studies have shown mixed findings for structural validity of the AQ-10 and AQ-50, and there is moderate evidence that both versions have satisfactory discriminant validity. There is a lack of data on content validity, and diagnostic validity also varies for both measures from an accuracy (AUC) of 0.65-0.99 depending on the study (Baghdadli et al., 2017). Wigham et al. (2019) were particularly concerned about the low specificity shown by the AQ-10 and AQ-50, cautioning that it should not be used to triage referrals. Both Conner et al. (2019) and Jones et al. (2021) found it was a poor predictor of diagnosis in clinical outpatient samples, including in the context of an NHS specialist autism service. In a sample from an autism diagnostic assessment clinic, Ashwood et al. (2016) found that AQ scores also did not predict diagnosis. The specificity was very low at 0.29 and a negative predictive value (NPV) of 0.36 – 64% of patients scoring below the cut-off of 32 were false negatives. The authors recommended a lower cut-off for screening ($\geqslant$26), which improved specificity. The AQ had adequate sensitivity and positive predictive value (PPV). Brugha et al. (2020) found the AQ to be satisfactory in a mental health patient population, with AUC=0.77, sensitivity 0.79 and specificity 0.77, and a moderate correlation between AQ scores and ADOS outcome.

Internal consistency for the AQ is mixed (Baghdadli et al., 2017; Nishiyama et al., 2014; Hudson et al., 2024). There is moderate to strong evidence for satisfactory test-retest reliability in the AQ (Baghdadli et al., 2017), which was also rated excellent by Brugha et al. (2020). Ashwood et al. (2016) found that for both the AQ-50 and AQ-10, there was good inter-rater reliability between self-report and informant-report versions. They suggest that this indicates that scores were not affected by issues of self-insight on the part of autistic adults.

Evidence for the validity of RAADS-R is also mixed. Baghdadli et al. (2017) reported satisfactory content, discriminant and convergent validity, all with moderate evidence to support these. There was unsatisfactory evidence for sensitivity and specificity, which ranged from 0.77-1.00 (Baghdadli et al., 2017). Diagnostic validity as indicated by AUC is heterogeneous. Conner et al. (2019) found it was a poor predictor of diagnosis in a clinical outpatient sample, AUC=0.58). Wigham et al. (2019) also reviewed RAADS-R and suggested the low specificity was again a concern. Brugha et al. (2020) found more promising results for RAADS-R in a mental health patient sample, with a sensitivity of 0.75 and specificity of 0.71, and moderate correlations between RAADS-R scores and ADOS outcomes. They did, however, suggest a higher cut-off threshold, as using the recommended one identified over half of patients as possibly autistic. Sturm et al. (2024) found that scores on all items of the RAADS-R were significantly different for those who were diagnosed autistic as opposed to non-autistic respondents. Those who were self-identified autistic scored significantly differently from those with a diagnosis on 4 items. They found a threshold of 35 to have the optimum sensitivity and specificity (88.31 and 85.65 respectively). Some studies have found satisfactory internal consistency and test-retest reliability for the RAADS-R (Baghdadli et al., 2017; Brugha et al., 2019; Sturm et al., 2024), though Baghdadli et al. (2017) noted that internal consistency was poorer for the language domain and that evidence for internal consistency overall was limited.

PPV for the M-CHAT appears relatively low across studies, ranging from 0.06-0.60 (Levy et al., 2020). Allison et al. (2021) reported that specificity for the M-CHAT in previous research was high, but it had low sensitivity and PPV. Charman and Gotham (2013) reported that PPVs for the M-CHAT show improvement if the test is administered in two stages, with those initially detected by the M-CHAT receiving follow-up interviews one month later as part of the screening procedure prior to assessment; however, this decreases sensitivity. A meta-analysis of the M-CHAT with follow-up (M-CHAT-R/F) found a pooled specificity of 98% and sensitivity of 78% (Santos et al., 2024). Wallis and Guthrie (2024) also found evidence to support the inclusion of M-CHAT-R/F in screening. Effectiveness of the M-CHAT has been found to differ across ages, settings and contexts (Levy et al., 2020), with its performance in clinical samples poorer than in research contexts (Wallis & Guthrie, 2024).

In a meta-analysis, the SCQ showed a wide range for sensitivity (0.47-0.96) and specificity (0.52-0.99). Findings on predictive validity also vary, resulting in an AUC of 0.827 in the meta-analysis. Its performance was best in representative population samples, and lower within clinical samples (Chesnut et al., 2019). A study comparing three screening instruments found the SCQ had the highest AUC compared to the SRS and CCC (see next section for details), at 0.90 (Charman et al., 2007), but a study of child and adolescent mental health patients in the community achieved an AUC of only 0.52, with sensitivity between 0.73-0.84 and specificity between 0.13-0.36. Parent and teacher scores were not significantly related, suggesting poor inter-rater reliability, and neither parent- nor teacher-reported scores were related to diagnostic outcome (Hollocks et al., 2019).

Baghdadli et al. (2017) found mixed evidence for the validity of the SRS-2; evidence for its sensitivity and specificity was rated unsatisfactory, and evidence for diagnostic validity was lacking. Sensitivity estimates range from 0.75-0.92 and specificity from 0.67-0.94 (Charman & Gotham, 2013; Bruni, 2014; Thabtah & Peebles, 2019). These appear to be higher for children than adults (Bruni, 2014). The SRS-2 shows low to moderate correlations with diagnostic outcomes (Bruni, 2014), though Charman et al. (2007) reported it showed no significant difference from the CCC. SRS-2 has been assessed as having satisfactory internal consistency and structural validity (Baghdadli et al., 2017; Bruni, 2014) but has been found to have low discriminant validity (Nishiyama et al., 2014). The total score has been reported as being more

reliable than the subscales alone (Bruni, 2014). Inter-rater reliability ranged from 0.61-0.92 but there is a lack of test-retest data available (Bruni, 2014).

Limited evidence for the ASBQ showed satisfactory structural validity and internal consistency, with only limited data available for its diagnostic accuracy and discriminant validity (Baghdadli et al., 2017). For the ASSQ, Kopp and Gillberg (2011) demonstrated sufficient discriminant validity between autistic and non-autistic participants with AUC 0.79. The ASSQ has been reported to have 90% accuracy using a cut-off threshold of 13, a sensitivity of 0.91 and specificity of 0.86 (Thabtah & Peebles, 2019). Studies testing the SACS-R in large samples of children have found it to have very high diagnostic validity, with sensitivity between 0.62-0.82 and specificity of 0.99 (Barbaro et al., 2022; Mozolic-Staunton et al., 2020).

Thabtah and Peebles (2019) and NHS England (2023a) note that updates to diagnostic criteria may affect the validity of screening tools designed prior to the changes. Wallis and Guthrie (2024) also note that a lack of follow-up with people who screen negative may mean tools have lower accuracy than is reported in research.

A.3.4 Acceptability and feasibility

The AQ-10 is recommended for screening of those suspected to be autistic in NICE guidelines, due its brief nature (Wigham et al., 2019). However, in 2021 concern was raised in the Lancet journal that NICE guidance on diagnosis of adults erroneously recommended a score of 7 or above as the threshold for a positive screen rather than 6 or above, which is the threshold that has been defined in research on this tool and approved by NICE itself (Waldren et al., 2021). NICE guidelines were amended in response to this, but it is possible some providers, such as GPs, could be using out-of-date guidance. Participants in Brugha et al.'s (2019) study completed both the AQ and RAADS-R, with higher rates of completion compared to RAADS-R indicated by fewer missing items. Several versions of the AQ are also easily available online and no specialist training is required to complete or score the AQ (Wigham et al., 2019). Ashwood et al. (2016) carried out a cost-benefit analysis for the amount saved using the AQ for screening versus the amount of autistic people who would go undiagnosed as a result. Based on their figures for diagnostic accuracy, this suggested "for every 10 patients denied a referral, £23,050 would have been saved but seven [autism] cases would go undiagnosed" (Ashwood et al., 2016, p.2601). Research has also found cultural differences in how parents interpreted items in the AQ, including comparisons to peers and cultural norms such as class, and differences in how specific words are understood in different languages (Genovesi et al., 2023).

Stacey and Cage (2022) asked autistic participants of a research study for feedback on use of the RAADS-14 (a shorter version of the RAADS-R) when used for research purposes, which may also provide insight as to its acceptability as a screener. A major issue on this measure was its response options that participants reported as finding difficult, feeling that a "sometimes" option was needed to record how their traits may fluctuate day to day or could be affected by issues of social context such as masking. Through analysis of item response curves on the RAADS-R, Sturm et al. (2024) found that only two response options were needed ("not true" and "true now and when I was younger"), with the options "true only now" and "true only when I was younger" providing no additional significant difference. Furthermore, the treatment of the response options as ordinal was not justified. Their findings suggest that this tool could be simplified by providing just two scoring options in a similar way to the AQ (which has four response options which are most often split into two scoring options). Other issues reported by Stacey and Cage's (2022) sample included feeling that the RAADS-14 questions lacked nuance and relevance, did not quantify what was meant by judgement-based terms such as "highly" and "extremely", and participants often wanted to be able to expand on their answers

with qualitative information. From a clinical perspective, the RAADS-R is also readily available and requires no training to use (Wigham et al., 2019).

In a case study report, the Autism Spectrum Rating Scale was chosen for assessment of needs for a child due to being brief and efficient to use compared to the more effective Gilliam Autism Rating Scale where a clinician's availability was limited (Livanis & Mouzakitis, 2010), highlighting how practical considerations can affect tool choice. Wigham et al. (2019) noted that developmental history and informant-report can be harder to obtain for adults where tools require these. In children with cerebral palsy, Påhlman et al. (2020) found that response rates and amount of completed items were lower for screening questionnaires of children with more severe levels of disability, in a sample with cerebral palsy using the ASSQ and SDQ.

A study found that scores on the self-reported adult SRS-2 varied depending on adjustments to the social reference group referred to in questionnaire items (the respondent's in-group or out-group; that is, asking how autistic people vs. neurotypical people viewed their behaviour), and this affected both autistic and non-autistic respondents' scores (Gernsbacher et al., 2017). Without specific instructions about social context, scores could be affected by respondents' assumptions about who they are expected to compare themselves to.

A review suggested that primary care in the United States typically uses the M-CHAT for screening in young children because it is widely available, brief, and has been translated into multiple languages (Wallis & Guthrie, 2024). They compared this to the SCQ, which requires a licence to use in practice and so is more rarely used for screening, but sometimes included in the lengthier diagnostic assessment process. They also reported that clinicians can face barriers administering screening tools, including reluctance to discuss autism with families, limited time, and limited knowledge about autism care pathways. Additionally, screening tools that have not been administered or completed correctly can cause delays. Services should have systems in place to deal with issues around delivery, data collection and communication of outcomes (Wallis & Guthrie, 2024).

Green et al. (2022) suggested digital apps accessible to parents could be used prior to diagnostic assessments to help parents track their child's outcomes over time. They proposed that this could reduce the demand on diagnostic services by giving parents support at the early stages and time to review progress before pursuing a referral. Apps such as ASDetect (Barbaro & Yaari, 2020), which give parents access to the SACS-R tool, could facilitate this and similar mechanisms could be put in place for adults. This would bridge the gap between services only conducting needs-based assessments using bespoke or unvalidated tools, while providing a system for managing service caseloads.

A.3.5 Consistency across populations, flagging of alternative/co-occurring conditions, and outcomes

Belcher et al. (2023) found that 41 AQ items were biased by gender between matched pairs, in particular items around social and communication issues. Murray et al. (2016) found no gender differences when looking at overall scores, indicating that gender biases on individual items were accounted for in the overall score (that is, bias of some questions towards males were cancelled out by questions biased towards females and vice versa). In Brugha et al.'s (2019) study, use of the AQ and RAADS-R identified that more women than men from their sample scored above the threshold for autism assessment. Some evidence suggests that the age of a child influences how parents interpreted items on the AQ (Genovesi et al., 2023). On the RAADS-R, Sturm et al. (2024) found some evidence for differential item functioning by age, gender and diagnosis. Ashwood et al. (2016) found that identification of generalised anxiety disorder predicted higher AQ-10 scores, potentially leading to false positives in those with

generalised anxiety disorder. Scores for the SRS-2 were correlated with traits of psychoticism and distress, and, while it showed good sensitivity between genders, its specificity for women was reduced compared to men (Wigham et al., 2019). Original research on the ASBQ found it was able to differentiate autism from depression, ADHD and schizophrenia (Horwitz et al., 2016). Mean scores on the ASSQ were significantly higher for autistic children compared to those with only ADHD, and girls' scores did not differ significantly from boys' except on one item (demand avoidance; Kopp & Gillberg, 2011). There is a lack of evidence for how these tools function in those who are transgender, non-binary or gender nonconforming.

The SRS-2 showed lower specificity in those more likely to have intellectual disabilities (Charman et al., 2007). However, in a clinical sample, the SRS-2 showed satisfactory internal consistency between gender, age and clinical subgroups (Bruni, 2014). Charman et al. (2007) found the SCQ, SRS and CCC all had lower specificity in patients who showed more behavioural problems. A review suggested higher thresholds for the SCQ are more reliable for ruling out other conditions (Thabtah & Peebles, 2019). The SCQ and CCC appeared to perform similarly in those with an intellectual disability, but SCQ scores differed by gender and country of administration (Charman et al., 2007). A review suggested the SCQ has consistently high sensitivity and internal consistency when used for people with an intellectual disability, but lower specificity (Metcalfe et al., 2020). However, the SCQ has also been adapted for those with an intellectual disability as the SCQ-AID (Derks et al., 2017). The M-CHAT had good sensitivity but lower specificity in children with Down syndrome. When combined with the SCQ, the M-CHAT led to false positives in people with an intellectual disability who also had visual and hearing disabilities (Metcalfe et al., 2020). These results only come from one study. Påhlman et al. (2020) found a sensitivity of 0.79 for the ASSQ in children with cerebral palsy; 35% of their sample screened positive for autism and 29% for both autism and ADHD, using the ASSQ and SDQ. Co-occurrence was also higher in people with a more severe presentation of cerebral palsy, people with an intellectual disability and people with epilepsy (Påhlman et al., 2020).

Summerill et al. (2024, pre-publication) reviewed 24 articles examining the use of tools and measures for autism screening in older adult groups. No studies had evaluated these tools' effectiveness in the diagnosis of older adults specifically. In samples which included older adults, measures such as the AQ and SRS did not account for age-related effects, and some tools such as the Gilliam Autism Rating Scale correlated with symptoms of dementia. The Broad Autism Phenotype Questionnaire has shown some limited evidence of being reliable in identification of older adults (Summerill et al., 2024).

Some tools, such as the M-CHAT, have been translated for accessibility and cultural relevance in low- and middle-income countries and in people from minoritised ethnic and cultural backgrounds, but further research is needed to ensure tools are adapted in this way and undergo rigorous testing (Wallis & Guthrie, 2024). Research suggests that disparities exist in diagnosis between ethnic groups, with mixed impacts including later recognition of autism in some ethnic minority groups, but higher rates of diagnosis in other groups compared to the ethnic majority of that country (Tromans et al., 2020; Wallis & Guthrie, 2024). Tromans et al. (2020) suggest that differing practices and policies across regions may impact these trends. Wallis and Guthrie (2024) report that patients from minority backgrounds are less likely to be referred for a full assessment even after a positive screen, and that quality improvement initiatives for services should encourage equitability towards marginalised groups. This suggests there is a high need for clinician judgement to be utilised in populations with high cultural diversity. Perera et al. (2017) developed a culturally sensitive picture-based screening tool, which showed a sensitivity of 0.88, specificity of 0.93, PPV of 0.95 and NPV of 0.84. Its

sensitivity was also good for discriminating between autism and other developmental disorders (0.88), with lower specificity (0.60), PPV of 0.78 and NPV of 0.77.

Evidence is lacking around how scores on questionnaires may be associated with outcomes for children and adults, which could help with triage. In a case study, the Autism Spectrum Rating Scale was used to identify areas of social skills to focus on in a child's educational setting, which led to apparent improvement in the desired social skills (Livanis & Mouzakitis, 2010). However, this was judged based on normative values such as what constitutes an acceptable conversation. In Chester & Wirral NHS Trust, a bespoke profiling tool, Thinking Patterns in Autism, is used in the assessment process to provide patients with a profile of strengths and needs for post-assessment support (Tollerfield & Pearce, 2020).

There has been a lack of attention in studies towards risk of harm from use of these instruments (Levy et al., 2020), although Hollocks et al. (2019) report that as autism assessment is non-invasive, there is relatively low risk of harm from false positives, suggesting that low specificity on a test is less important to prioritise than sensitivity. Wallis and Guthrie (2024) also note that false positives from screening in children contain a high proportion of children who go on to be diagnosed with other conditions, suggesting screening for autism may help to flag other support needs. However, false positives can cause harm on a wider scale by extending waiting lists, meaning longer waits for support both for those who eventually receive an autism diagnosis and those for whom other diagnoses would be more appropriate. This can also have cost implications for services as well as impacting the quality of life of families and individuals.

A.3.6 Tools for broader neurodevelopmental assessment or assessment of needs

The A&SQ is a commonly used parental-report assessment for ages 1-66 months, comprising communication, motor skills, problem-solving and personal-adaptive skills (Muthusamy et al., 2022). The DAWBA is an informant-completed interview that aims to assist with diagnosis of a range of conditions (Coscini et al. 2020). A self-report interview is also included for completion by children aged over 11 years. The CCC (Bishop, 1998) is a 70-item scale covering language and communication in children, scored on a three-point scale, with 9 subscales. The SDQ (Goodman, 1999) is a 25-item measure of emotional and behavioural traits on 5 subscales: "emotional symptoms, conduct problems, hyperactivity/ inattention, peer relationships, and prosocial behavior" (Russell et al., 2013). It was developed for parental-report of children, but self-report versions are available for adults and children aged 11-17. The Early Years Foundation Stage Profile is "a universal educational assessment […] conducted on all children in their first year of schooling in the UK" by teachers, covering developmental domains relevant to autism (Wright et al., 2019, p.2). The Developmental Behaviour Checklist Autism Screening Algorithm is an informant-rated measure comprising 29 items for children between age 4-18 (Brereton et al., 2002). An early-screen version has also been developed for children between 20-51 months (Gray et al., 2008). The Conners Ratings Scales are widely used in assessment for ADHD but can also assess a variety of issues, with self-report, parent- and teacher- report and adult-report versions available (Collett et al., 2003). The BRIEF (Gioia et al., 2000) assesses executive functioning in children aged 11-18 including informant- and self-report forms.

Muthusamy et al. (2022) identified 43 studies evaluating the A&SQ, 77% of which were rated low for risk of bias in patient selection, with higher risk of bias found for index tests. When used to diagnose any developmental delay, the A&SQ showed pooled sensitivity of 0.77, specificity of 0.81 and AUC 0.86; higher accuracy was found in older children compared to younger children (Muthusamy et al., 2022). In a large sample of toddlers, Hardy et al. (2015) found 87%

agreement between communication scores on the AS&Q with a positive M-CHAT-R outcome, and 95% agreement with an autism diagnosis. This questionnaire is brief with a 10–15-minute completion time and is inexpensive for services.

A limited number of studies on DAWBA showed mixed findings for its validity as a screening tool for autism, with one study demonstrating high sensitivity but low specificity in a sample of children referred for autism assessment, and a non-significant AUC (Coscini et al., 2020). Reducing the cut-off improved sensitivity moderately but reduced specificity, and inter-rater reliability was also poor (Coscini et al., 2020). A twin study found both high sensitivity and specificity for a version of DAWBA using a computerised algorithm to assess scores against diagnostic outcomes; there was only minimal difference when scored by a clinician (McEwen et al., 2016). PPV and NPV were also high, and DAWBA's performance was improved when used in conjunction with the ADOS (McEwen et al., 2016).

A study found the CCC to have high sensitivity but low specificity for detecting autism, with a similar performance across intellectual functioning levels but reduced specificity in those with behavioural problems (Charman et al., 2007). A study showed the SDQ to have a high rate of false positives for autism and ADHD, with the sensitivity for autism at 0.79 and specificity of 0.93. On one model tested, the sensitivity for ADHD was only 0.30. There was also low inter-rater reliability between parent and teacher scores, especially for the emotional subscale; the researchers found that autism was associated with higher scores on emotional symptoms, which they suggest may be linked to the overlap between autism and anxiety and depression (Russell et al., 2013). Russell et al. (2013) noted that questionnaires were most frequently completed by mothers of children, representing 96.7% of their sample in the Millennium cohort. A review found that the parental report version of the SDQ had highest specificity for screening for psychiatric conditions, and that the SDQ performed consistently across settings, but lacked data on differences across demographics (Warnick et al., 2008). A pilot study found the SDQ may have value in triage for adults, by indicating potential co-occurring conditions, with acceptable validity and reliability but low specificity. There was some overlap between subscales, and parental-report was found to have higher validity than self-report (Findon et al., 2016). However, a study of adolescents found the SDQ's factor structure was an inadequate fit for the autistic group and suggested it required revising to be suitable for indicating co-occurring mental health needs in this group (Turcan et al., 2024).

A review suggested the BRIEF has been found to have high internal consistency for scales with a larger number of items but a range of test-retest correlations. Inter-rater correlations between adolescents and their parents have been reported as moderate to high, and lower between adolescents and their teachers, but still significant. Findings did suggest scores could be affected by age and gender (Walker & D'Amato, 2006); t-score tables can be used to convert results into a standardised score accounting for factors such as age, so should be used for scoring.

A review found the DBC to have a sensitivity of 0.83 and specificity of 0.48 for screening for autism, with a cut-off of 17 producing optimum results (Thabtah & Peebles, 2019). A study found that this scale may be less effective for those with behavioural problems (Witwer & Lecavalier, 2002). The Conners rating scales tend to have moderate to high internal consistency, but lower test-retest scores, and variations in specificity and sensitivity between different versions of the scale, with the parent and self-report versions showing highest values (Collett et al., 2003). The adult version of the scale showed low internal reliability on some subscales, potential overlap with other conditions (as measured by convergent and discriminant validity), and potential gender bias (Smyth & Meier, 2019). Most research on this tool has been carried out regarding screening for ADHD.

Wright et al. (2019) found that children with a low Early Year's Foundation Stage Profile score were approximately 50 times more likely to have an autism diagnosis recorded. The researchers suggested that the tool may have use as a population screener (although note that this is not recommended for autism), or scores could be considered case-by-case for individual referrals to identify likelihood of autism. As this tool is already used routinely in schools, feasibility is high for its use in screening. In children identified through this method, males, white British children, and those of higher socio-economic status were groups most likely to have an autism diagnosis.

In Scotland, the Fife Neurodevelopmental Questionnaire (Muggleton, 2024) has been developed as a transdiagnostic tool (i.e., a tool that is designed to aid with understanding an individual's needs, irrespective of diagnosis). This tool has been developed by an autistic clinical psychologist, along with colleagues. The Fife team argue this type of tool may be more acceptable to parents because it collects information about the child's history at one single point, rather than separate tools being needed to assess likelihood of different neurodevelopmental conditions (Autism Network Scotland, 2020). However, it has not been developed for the purpose of screening or diagnosis.

Appendix 6. Self-assessment questions

Validity and reliability

- What was the original purpose of the tool(s) in question?
- Has there been research that has carried out psychometric validation of the tool(s)?
- What aspects of validity and reliability have been studied? What is missing?
- Have the findings been peer-reviewed?
- Has the tool been tested in settings and populations that are representative of the service/setting?
- How much exposure have the population had to this tool already? Will prior knowledge affect its reliability?

Acceptability

- Is there any research available that explores the acceptability of the tools for different user groups?
- Does its acceptability to patients align with its acceptability to professionals involved in screening and triage in the autism assessment pathway?
- How burdensome is the tool to use – for patients, for professionals and for other users?
- What difficulties often arise when completing or scoring the tool?
- What aspects of the tool might indicate its acceptability for the different user groups? How can this be assessed?
- Are there ways to mitigate for low levels of acceptability for a tool that is otherwise well-evidenced? For example, would it help to give clearer information about why certain terminology is used?

Feasibility and efficiency

- Has the cost-effectiveness of the tool(s) in question been evaluated in the appropriate context? How long ago was cost-effectiveness evaluated?
- Is there any evidence that a tool may be more effectively used in combination with others?
- How much clinician time is required to use the tool? Is clinician time required to support patients with completing tools at referral? How much clinician time is required to review a tool following a referral, or for triage?
- Does the tool require a licence to use?
- Do clinicians have the appropriate technology and training to support using the tool in its intended way?
- Can communication with service users be improved to ensure tools are completed in their intended way (such as explaining why a tool is scored a certain way)?
- Are tools presented in accessible formats?

Extent of evidence

- Has the tool been included in any meta-analyses or systematic reviews? If so, how many reviews, and how many studies of the tool have been identified?

- How long ago were meta-analyses/reviews carried out? Is there additional research that has been produced since?
- Are findings on the tool consistent or do they vary? On what aspects do they vary?
- What do reviews and individual studies indicate about the use cases the tools have been applied to?
- Who has carried out studies on the tool? (for example, if all evaluations come from the team who designed the tool, there is a high risk of bias)

Design characteristics of studies

How do previous studies on the tool(s) compare on:

- Methods used to verify diagnosis (for example, self-report of an existing diagnosis; diagnostic assessment by a clinician; use of another tool as a reference test)?
- Use of reference tests – Which tools have been used as the reference test? Are they well-validated? Were all participants given the same tests? Were reference and index tests rated without knowledge of results on the other test?
- Sampling – What size samples have been used? How were samples recruited? Was any randomisation used? Do the inclusion/exclusion criteria represent people using services?
- Design of procedures – Have case-control designs been avoided (that is, comparing scores of autistic people with scores of non-autistic people, which would be required to determine whether the tool is successful at identifying autism)? Has there been follow-up to assess any change over time?

Outcomes

- What type of output does the tool produce? Who is it useful for? How does it help to inform next steps?
- What information is not captured by the tool that may be crucial to guiding clinical judgment?

Flagging of alternative or co-occurring conditions

- What co-occurring conditions need to be identified or ruled out in the population in question?
- What does research suggest about how well the tool can differentiate possible autism from other conditions?
- What does research suggest about how well the tool can indicate the presence of conditions that co-occur with autism? How does this affect triage, outcomes and support?
- If the tool assesses broader neurodevelopmental traits (that is, it is not specific to autism; for example, the SDQ or CCC), has it been validated for screening or triage in autism assessment?

Consistency across populations

- Has previous research investigated whether the tool's performance differs by age, gender, ethnicity and culture or disability?

- What aspects of the service's screening and triage procedures affect access or outcomes for people from marginalised or underserved groups? What is the service doing to ensure processes are fair and inclusive? Does this have any influence on which tools or procedures can be used?