Review

Psikiyatride Güncel Yaklaşımlar-Current Approaches in Psychiatry 2026; 18(2):510-528

doi: 10.18863/pgy.1636985

Psychometric Challenges in the Screening and Diagnosis of Autism Spectrum Disorder: A Review for Special Education and Guidance Services

Otizm Spektrum Bozukluğunun Tarama ve Tanılama Sürecinde Psikometrik Zorluklar: Rehberlik ve Özel Eğitim Hizmetlerine Yönelik Bir Derleme

Eylül Akar[1], Adem Kantar[2], Berfu Ayan[2]

[1]Kastamonu University, Kastamonu
[2]Erzurum Technical University, Erzurum

ABSTRACT

Early diagnosis of autism spectrum disorder and planning interventions appropriate to individual needs are essential for the development of individuals. In this process, valid and reliable assessment tools are critical for diagnosis accuracy and the effectiveness of interventions. This review aims to present information on good practice for selecting assessment tools, preparation for the data collection process, data collection, analysis, and reporting stages to support valid and reliable screening and diagnosis processes in evaluating autism spectrum disorder. In this context, assessment tools used in the screening and diagnosis processes of autism spectrum disorder were examined in the context of validity and reliability, the criteria to be considered in selecting these tools were emphasized, and measurement errors encountered in the application processes were focused on. Suggestions are presented to minimize the impact of measurement errors. This study is thought to guide practitioners involved in the evaluation process of autism spectrum disorder.

Keywords: Autism spectrum disorder, screening, diagnosis, assessment tool, validity, reliability

ÖZ

Otizm spektrum bozukluğunun erken tanılanması ve bireysel ihtiyaçlara uygun müdahalelerin planlanması, bireylerin gelişimi açısından önem taşımaktadır. Bu süreçte, tanının doğruluğu ve müdahalelerin etkililiği için geçerli ve güvenilir değerlendirme araçlarının kullanımı kritik öneme sahiptir. Bu derlemede, otizm spektrum bozukluğunun değerlendirilmesinde geçerli ve güvenilir tarama ve tanılama süreçlerini desteklemek amacıyla, değerlendirme araçlarının seçimi, veri toplama sürecine hazırlık, veri toplama, analiz ve raporlama aşamaları için iyi uygulamaya ilişkin bilgileri sunmak amaçlanmaktadır. Bu doğrultuda, otizm spektrum bozukluğunun tarama ve tanılama süreçlerinde kullanılan değerlendirme araçları geçerlik ve güvenirlik bağlamında incelenmiş, bu araçların seçiminde dikkat edilmesi gereken kriterler vurgulanmış ve uygulama süreçlerinde karşılaşılan ölçme hatalarına odaklanılmıştır. Ölçme hatalarının etkisini en aza indirmek için öneriler sunulmuştur. Bu çalışmanın otizm spektrum bozukluğunun değerlendirme sürecinde yer alan uygulayıcılara yol gösterici olacağı düşünülmektedir.

Anahtar sözcükler: Otizm spektrum bozukluğu, tarama, tanılama, değerlendirme aracı, geçerlik, güvenirlik

Address for Correspondence: Eylül Akar, Kastamonu University Faculty of EducationDepartment of Special Education, Kastamonu, Türkiye
e-mail: eylulakar@kastamonu.edu.tr
Received: 10.02.2025 | **Accepted:** 23.07.2025

Introduction

Autism Spectrum Disorder (ASD) is a lifelong neurodevelopmental disorder characterized by difficulties in several areas such as limitations in social interaction and communication skills, repetitive/restricted behaviors, adherence to routines, and insistence on sameness (American Psychiatric Association [APA] 2013). Since individuals with ASD are evaluated across a broad spectrum, the severity and symptoms of ASD can vary from person to person. This means that each individual can be affected by ASD characteristics differently and to varying degrees of intensity. Individuals with ASD generally exhibit behaviors such as difficulty with gestures, facial expressions, and maintaining two-way conversations, limitations in the use of body language, inability to make eye contact, limited ability to engage in symbolic play, and inability to maintain continuity in relationships with the environment, which are classified as social communication and interaction problems; They may exhibit behaviors classified as repetitive/obsessive behaviors and limited interest and activity, such as excessive interest in particular objects, excessive adherence to routines, sudden reactions to changes in routine, repetitive/stereotyped motor movements, and under- or over-reaction to sensory stimuli when necessary (APA 2013).

According to the latest report published by the Centers for Disease Control and Prevention (CDC), one in every 31 children worldwide has ASD (CDC 2025). While the prevalence rate of ASD varies by country, no report has yet been published in Türkiye (Köse et al. 2017). With this significant increase in ASD in recent years, awareness of ASD has increased, and screening efforts have gained importance. The American Academy of Pediatrics recommends screening children for ASD risk, especially between 18 and 24 months of age (Johnson et al. 2017). The Ministry of Family and Social Policies' (Aile ve Sosyal Politikalar Bakanlığı 2023) Second National Action Plan for Individuals with Autism Spectrum Disorder aims to strengthen diagnosis, follow-up, and intervention programs and to ensure data-based monitoring by generating statistics. Evaluation of individuals with ASD includes clinical interviews with parents or primary caregivers, medical, psychological, and educational records review, direct play, and cognitive and developmental evaluation (Kılınç et al. 2019).

Research shows that early interventions, particularly those offered in the first three years of life, play a critical role in the development of children with ASD (Corsello 2005, Dawson et al. 2010, French and Kennedy 2018), and that specific interventions related to social communication and language skills affect the development of these children and the course of ASD (Rogers 2005). Therefore, accurate assessment is crucial for early recognition of ASD and providing interventions tailored to the individual's needs (Selimoğlu et al. 2013). In other words, early diagnosis allows children with ASD to receive appropriate interventions. Therefore, selecting tools used in the assessment process is crucial in influencing this process. There are review studies in the literature examining the tools used in the assessment of ASD (Kılınç et al. 2019, Yazıcı and Akman 2020). These studies provide information on the current status of assessment tools used in the screening, diagnosis, and educational assessment of ASD. This review aims to provide practitioners and researchers with recommendations regarding potential errors encountered in the psychological measurement and assessment process during the screening and diagnosis phases of ASD, their sources, and how to reduce these errors. To this end, the importance of psychological assessment in ASD, a brief introduction and function of assessment tools used in screening and diagnosis for ASD, considerations for selecting assessment tools, and recommendations for practitioners are provided.

Significance of Psychological Assessment in ASD

Research on the diagnostic and screening processes for ASD has focused particularly on the importance of psychological assessment and the effectiveness of various assessment tools used in this process (Selimoğlu et al. 2013, Aksoy and Şahin 2017, Kılınç et al. 2019). Changes to the ASD diagnostic criteria in the Diagnostic and Statistical Manual of Mental Disorders (DSM), particularly with the transition from DSM-IV to DSM-5, have revealed significant differences in the ASD diagnostic criteria. These changes are based on research conducted to understand ASD better. Genetic, neurodevelopmental, and behavioral studies have demonstrated that ASD should be considered a spectrum disorder (Huerta et al. 2012). Because overlap

and ambiguity were observed among the subcategories in DSM-IV, the single diagnostic category "autism spectrum" aimed to reduce these differences and ambiguities. The spectrum concept was adopted by expanding the diagnostic criteria in DSM-5 (APA 1994, 2013).

In the DSM-IV, a pervasive developmental disorder (PDD) diagnosis was made if the child's symptoms met specific diagnostic criteria, but the severity of the symptoms was not graded. In the DSM-5, severity levels were added to the diagnosis of ASD, based on the amount of support an individual requires: "individuals needing little support, individuals needing moderate support, and individuals needing intensive support" (APA 1994, 2013). This grading system allows for tailored interventions to the needs of individuals with ASD. These changes have created new challenges and adaptation requirements for fieldworkers, but they have also made significant strides toward increasing the accuracy and comprehensiveness of the diagnostic process.

Early recognition and accurate diagnosis of ASD is crucial for initiating early educational interventions. Indeed, education provided during early childhood, while brain development continues, is known to yield more lasting and positive effects. This period is when the brain exhibits the highest rate of neuroplasticity (Hensch 2016). Furthermore, plasticity increases during critical periods when the brain is more sensitive to stimuli to acquire specific functions (Hensch 2016). Research indicates that children with ASD who begin education at an early age experience significant improvements in their social development, communication, language, speech, play, and cognitive skills (Landa 2008). Furthermore, early education enhances children's learning abilities and ability to adapt to challenging and stressful situations later in life (Turhan and Özbay 2016).

Because early intervention is possible through early detection, it is crucial to appropriately select screening and diagnostic tools and use them carefully and effectively (Elder et al. 2016). For early intervention programs to be planned and implemented, experienced screening personnel must determine the appropriate assessment tools, identify children needing early intervention, and ensure accurate diagnosis (Kurnaz-Adıbatmaz and Özyürek 2019). Professionals should be mindful of the possibility of biased responses, particularly in assessments based on information gathered from caregivers. In summary, both the validity and reliability of standard and non-standard tools used in the screening and diagnosis of ASD, as well as the experience of professionals, are critical to the success of early intervention programs.

Psychological Assessment Tools Commonly Used in the Screening and Diagnosis of ASD

The evaluation of children with ASD begins with the screening process. The screening process consists of a level 1 and level 2 screening model. In the level 1 screening model, all children are assessed for overall development, starting in the first months. In the level 1 screening model, children are evaluated for ASD at 9, 18, 24, and 30 months, particularly at 18 and 24 (Diken et al. 2012, Hyman Levy and Myers 2020). After the screening process, children exhibiting symptoms of ASD are taken into the diagnostic process for further examination and to confirm the ASD diagnosis. At this stage, children are evaluated according to the diagnostic criteria of the DSM-5 and the International Classification of Diseases, 11th Revision (ICD-11) (APA 2013, World Health Organization [WHO] 2024).

Screening Tools

This section describes level 1 and level 2 screening tools. Additionally, assessment tools with validated and reliable studies in Türkiye are examined for their psychometric properties. Tools used in screening ASD in national and international literature are presented in Figure 1.

Level 1 Screening Tools

They are quick and practical tools to identify children exhibiting early symptoms in a large population (Matson et al. 2010). These tools can be used by both healthcare professionals and parents to screen for

children at risk for developmental delay or ASD (Dumont-Mathieu and Fein 2005). The Denver Developmental Screening Test II, the Ankara Developmental Screening Inventory, the Gazi Early Childhood Assessment Tool, the Ages and Stages Questionnaire, and the Three-Item Direct Observation Screen are among the most commonly preferred level 1 screening tools for screening for developmental delay or ASD symptoms in early childhood (Eaves and Milner 1993, Rutter et al. 2003).

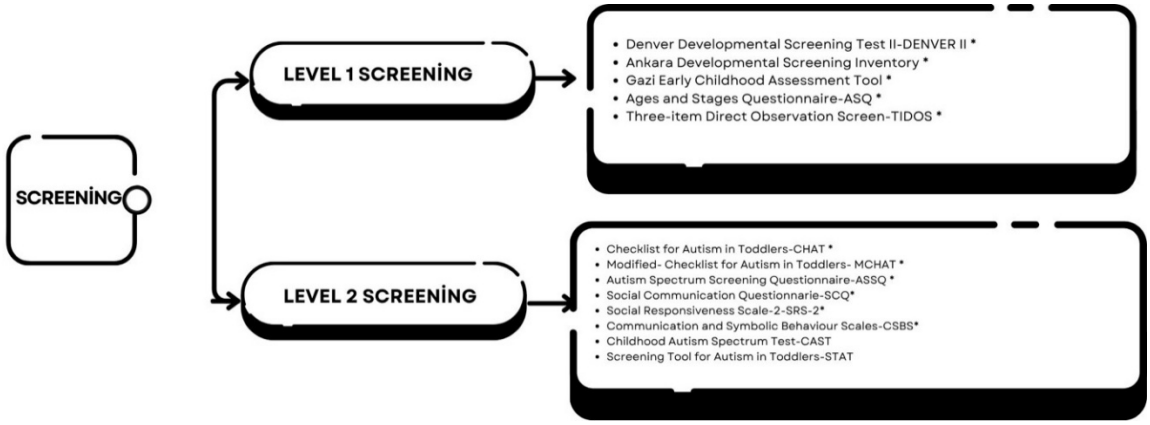

Figure 1. Screening tools in autism spectrum disorders

Assessment tools marked with an asterisk (*) indicate those that have Turkish validity and reliability studies (Günden and Diken, 2022; Uran-Kurtgöz, 2023).

Denver Developmental Screening Test II (DDST II)

The Denver Developmental Screening Test-II is a 125-item screening tool developed to identify developmental problems in children aged 0-6. It assesses four developmental domains: personal-social, fine motor-adaptive, language, and gross motor skills. The test was developed in 1967 and updated in 1990 (Frankenburg and Dodds 1967, Frankenburg et al. 1990). Frankenburg and colleagues (1992) subsequently re-evaluated the test in a larger population to address concerns about the test. The Denver II, standardized by Yalazet al. (2011), has a test-retest reliability of .86 and an inter-rater reliability of .90.

Ankara Developmental Screening Inventory

The scale was developed in Türkiye by Savaşır and colleagues (1994) to be administered to children with low socioeconomic status, considering the impact of environmental conditions on children's cognitive and psychosocial development. It aims to screen children's general development. The researchers assumed that mothers observe their children more frequently and understand them better. Therefore, they designed the scale so that mothers could respond with "yes" (1), "no" (0), or "I do not know" (0). It comprises 154 items with language-cognitive, fine-gross motor, and social skills-self-care dimensions. The reliability coefficients for the dimensions were observed to be adequate for all age groups (α >.84) (Savaşır et al. 1994).

Gazi Early Childhood Assessment Tool

It aims to evaluate the development of Turkish children who are not diagnosed with special needs, to organize their educational lives, and to guide them for more comprehensive evaluations by making a preliminary diagnosis of developmental disabilities (Işıl et al. 2015). Measurement is done by observing the child through developmental games, administering the test, and obtaining information from their parents when necessary (Temel et al. 2004). The Gazi Early Childhood Development Assessment Tool, developed by Temel and colleagues (2004), aims to measure the development of the 0-6 age group with four subtests: psychomotor, cognitive, language, and socio-emotional development, and 249 items (Baykan et al. 2002). It is scored by the examiners as yes (1) or no (0). Reliability coefficients for the subtests were calculated as α>.82, and inter-examiner reliability coefficients were calculated as α>.77, indicating a sufficient level (Temel et al. 2004).

Ages and Stages Questionnaire (ASQ)

The scale, edited and developed by Squires and Bricker at the University of Oregon, was first published as the "Infant/Child Monitoring Questionnaire" (Bricker et al. 1988). In 1999, it was renamed the "Ages and States Questionnaire (ASQ-2)" by Bricker and colleagues, and it took its final form following norm studies (Şişman 2021). The ASQ-2 is a developmental assessment scale that should be completed by children aged 3-72 months, their parents, caregivers, and the child's teacher. Bricker and colleagues (1999) observed that the test's stability and the correlation coefficient between raters (professional and caregiver) were .94. In this scale, sensitivity refers to the degree to which it identifies children with developmental disabilities, while specificity refers to the degree to which it identifies children with typical developmental courses. The ASQ's sensitivity level is .72, and its specificity level is .86. The fact that the ASQ can be used as a developmental screening inventory in different cultures (France, Spain, Denmark, Norway, China, Korea, Türkiye) suggests that cultural influences have been minimized (Kapçı et al. 2010, Yazıcı and Akman 2020). The validity and reliability of the ASQ-2 were tested on a Turkish sample by Kapçı et al. (2010) and compared with the Denver-II and Gazi Early Childhood Development Assessment Tool (GEÇDA). The ASQ consists of 30 items, measured monthly for 19 months, prepared according to children's age ranges. These questions test five subscales: communication, gross and fine motor skills, problem solving, and personal and social development. The overall reliability coefficient of the scale's dimensions (α>.82) is adequate (Kapçı et al. 2010).

Three-Item Direct Observation Screening (TIDOS)

The Three-Item Direct Observation Screening Test, developed by Öner and colleagues (2014), focuses on three parameters (responsiveness to name, joint attention, eye contact) in an observational manner. When this test is administered based on direct observation, a score of 1 in any parameter indicates that the individual is at risk for ASD. The child is observed for 15 minutes using several toys to observe skills such as responsiveness to name, joint attention, and eye contact. Responsiveness to name is scored 0-1-2, joint attention 0-1-2, and eye contact 0-1. In the study, sensitivity for ASD diagnosis was calculated as .73, and specificity as .70. In the direct observation test, sensitivity was calculated as .82 and specificity as .90 for the joint attention item; sensitivity was calculated as .89 and specificity as .91 for the eye contact item; and sensitivity was calculated as .67 and specificity as .87 for the name sensitivity item (Öner et al. 2014).

Level 2 Screening Tools

Level 2 screenings are explicitly conducted for children who exhibit symptoms of ASD or are considered at risk (Zwaigenbaum et al. 2015). These screenings take into account factors such as the child's developmental disability, family or teacher reporting behaviors that could be considered symptoms of ASD, or the presence of a family member previously diagnosed with ASD (Robins et al. 2014). The tools used for this purpose include ASD-related measurement items and serve to assess children in general for ASD (Charman and Baird 2002). In Türkiye, level 2 screening tools include those whose validity and reliability studies have been conducted with a limited number of clinically relevant participants, such as the Checklist for Autism in Toddlers (CHAT, Tetik-Kabil 2005) and the Modified Checklist for Autism in Toddlers (M-CHAT, Yıkgeç 2005).

Checklist for Autism in Toddlers (CHAT)

The first CHAT studies were conducted by Baron-Cohen and colleagues (1992) with two separate groups of 91 children to predict whether children who were unable to demonstrate joint attention and pretend play would be at risk of later being diagnosed with ASD. The scale's administration age was specified as 18 months. Furthermore, administering the CHAT to younger children was not recommended due to the increased risk of false positives. Subsequently, a second study was conducted with 16.235 18-month-old children (Baron-Cohen et al. 2000). In the original research, the items indicating high-risk criteria had a sensitivity of .18 and a specificity of 1.0. Positive prediction was calculated as .75 and negative prediction as .99. For all pervasive developmental disorders, the sum of the moderate- and high-risk items had a sensitivity of 0.21, a specificity of 0.99, and a positive prediction of 0.59 (Baron-Cohen et al. 2000). Additionally, Baron-Cohen et al. (2000) reported that the CHAT has high specificity for all pervasive

developmental disorders but low sensitivity for ASD. A validity study in Türkiye was conducted by Tetik-Kabil (2005) with 80 typically developing children, children with Down syndrome, children diagnosed with ASD, and children at risk for ASD. The test was found to accurately diagnose ASD (.86). In another study conducted by Hergüner and Özbaran (2010), criterion validity was examined by comparing the M-CHAT with the CHAT and was found to be adequate (.75).

Modified Checklist for Autism in Toddlers (M-CHAT)

Robins and colleagues (2001) developed the M-CHAT by updating and revising the CHAT. Aimed at screening toddlers aged 16-30 months for ASD, the M-CHAT is a 23-item screening tool that parents or caregivers can answer with a yes or no question. Children identified as high-risk with this screening tool are referred for more comprehensive assessments (Robins et al. 2001). The format and first nine items of the M-CHAT were taken directly from the CHAT. The observational section of the CHAT was not included. Some items in the scale directly test ASD, while others aim to reduce parents' anxiety and prevent them from focusing their attention on ASD (Kılınç et al. 2019). Six critical items of the scale (2, 7, 9, 13, 14, and 15) are directly related to ASD. The reliability coefficient for the six critical items in the original scale was .83, and the reliability coefficient for 23 items was .85. The scale was adapted to Turkish by Yıkgeç (2005). It was stated that item 6 should be included among the six critical items in the scale to identify seven crucial items. Predictive values could not be calculated due to the small number of cases in the sample. Kara (2014) tested the scale's validity, reporting predictive values as .98 for sensitivity, .89 for selectivity, .75 for positive prediction, and .99 for negative prediction. It was also demonstrated that this test yielded more accurate results when completed by experts rather than directly by parents. If the answer in bold is chosen in two or more of the questions 2, 5, 7, 9, 10, 14, 15, 17, 19, and 21 of the scale or in three or more of the 23 questions, it is revealed that the child is at risk of ASD (Yazıcı and Akman 2020).

Autism Spectrum Screening Questionnaire (ASSQ)

Ehlers et al. (1999) used the scale with children aged 6-17 years diagnosed with PDD who were of normal intelligence or mild intellectual disability (IQ≥50). The ASSQ, consisting of 27 questions designed to assess social interaction and communication difficulties, restricted and repetitive behaviors, and related challenges such as motor Clumsiness and Other Associated Symptoms, was reported to have a sensitivity of 62% and a specificity of 90% in screening high-functioning individuals (Ehlers et al. 1999). The revised original scale had adequate inter-rater reliability (r = .77) and test-retest reliability (r = .94) for ASD. Köse et al. (2017) calculated the test-retest reliability as .98 for the Turkish version of the scale, indicating an excellent level. When the scale's cut-off score was 16, diagnostic accuracy was calculated as .91, sensitivity as .94, and specificity as .89. The positive predictive value was .81, and the negative predictive value was .97 (Köse et al. 2017). Furthermore, researchers demonstrated through ROC analysis that a cut-off score of 26 was necessary to distinguish ASD cases from ADHD and OCD diagnoses. With a cut-off score of 26, the scale's overall diagnostic accuracy was .77, sensitivity was .63, specificity was .84, positive predictive value was .63, and negative predictive value was .84 (Köse et al. 2017).

Social Communication Questionnaire (SCQ)

The scale was developed to assess and screen for social interaction, communication, and restricted, repetitive, and stereotyped behaviors in individuals with a mental age of 2 and above and a chronological age of 4 and above (Berument et al. 1999, Rutter et al. 2003). The 40-item scale, answered yes or no, was designed based on parental report (Rutter et al. 2003). The internal consistency coefficient of the scale was reported to be .90, and specificity and sensitivity rates were above .80. The Turkish adaptation, validity, and reliability studies of the scale were conducted by Avcil et al. The scale's sensitivity for ASD diagnosis was calculated as 100%, and the specificity was .33 (Avcil et al. 2015).

Social Responsiveness Scale-2 (SRS-2)

The Social Responsiveness Scale-2 was developed to assess ASD symptoms in the context of social communication (22 items), social cognition (12 items), social awareness (8 items), social motivation (11 items), and restricted interests and repetitive behaviors (12 items). Parents or teachers can complete the scale. The scale has four different forms: preschool (2.5-4.5 years), school-age (4-18 years), adult (19 years

and older), and adult self-assessment form, and is scored using a four-point Likert scale. The scale demonstrated high reliability in the original study, with Cronbach's α internal consistency coefficient ranging from .93 to .95 (Constantino and Gruber 2012). The Turkish adaptation and validity-reliability study of the preschool form of the scale was conducted by Bakkaloğlu et al. (2021). The scale's item-total correlations were significant, and the factor structure was consistent with the original form. The internal consistency coefficient for the Turkish version was .95. The fit values for the scale were found to be at an acceptable level according to factor analysis (X2 = 5302.684, SD = 2005, p>.05, X2/SD = 2.64, RMSEA = .071, CFI = .78). These findings demonstrate that the scale is a practical assessment tool in both its original form and its Turkish version.

Communication and Symbolic Behavior Scales

The Communication and Symbolic Behavior Scale is a 22-question instrument explicitly developed to assess children's early childhood communication skills, symbolic behaviors, and social interactions (Wetherby and Prizant 2002). The scale is suitable for infants with a functional communication age of 6-24 months and a chronological age of 5-6 months, with a developmental age of less than 24 months. It can be administered using observations and parent reports. The scale assesses seven basic communication and symbolic skills (communication functions, gestural communication, vocal communication, reciprocity, social-affecting signaling, and symbolic behavior). In the original study, Cronbach's α internal consistency coefficient ranged from .80 to .95 (Wetherby and Prizant 2002). The validity and reliability studies of the scale for the Turkish language were conducted by Korkmaz (2009). The scale provides an interviewer-assisted assessment of seven primary caregiver domains using three components (infant checklist, caregiver questionnaire, and behavioral sample). The test-retest correlation coefficient for the scale ranges from .95 to .99 for all subcomponents. Furthermore, scale scores correlate significantly with Denver Developmental Screening Test language scores and Ankara Developmental Screening Inventory language scores (Korkmaz 2009).

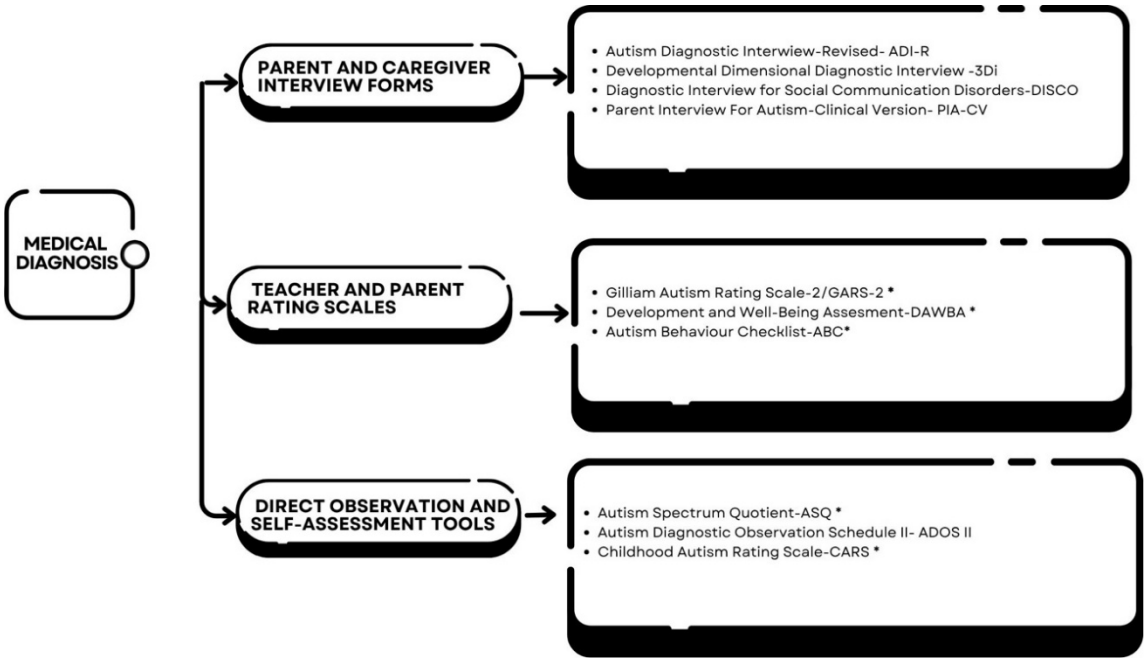

Figure 2. Diagnostic tools used in autism spectrum disorders

* Assessment tools marked with an asterisk (*) indicate those that have Turkish validity and reliability studies (Günden and Diken 2022, Uran-Kurtgöz, 2023).

Diagnostic Tools

Assessment tools used in diagnosing ASD are crucial in standardizing the diagnostic process and achieving reliable results. Diagnostic tools for ASD are based on two primary sources of information. The first is behavioral information and developmental history obtained from caregivers, while the second is direct

observation of behavior (İncekaş-Gassaloğlu et al. 2016). Among these tools, scales such as the Gilliam Autistic Disorder Rating Scale-2, the Autism Behavior Checklist (ABC), and the Childhood Autism Rating Scale (CARS) stand out. The Tools used to diagnose ASD in national and international literature are presented in Figure 2.

Teacher and Parent Rating Scales

Gilliam Autism Rating Scale-II (GARS-II)

The scale assesses ASD symptoms or diagnostic features in individuals between 3 and 23. Developed by Gilliam (1995), it was adapted to Turkish by Diken et al. (2012). The scale, consisting of 42 items and three subscales: social interaction (14 items), communication (14 items), and stereotyped behaviors (14 items), is scored using a four-point Likert-type scale. Sensitivity and specificity for ASD are 1 and 0.87, respectively, using a cut-off score 85. A score of 85 or above on the GARS-2 indicates a high risk for ASD, while a score of 69 or below indicates a low probability of ASD. The reliability coefficient of the adapted scale ranges from .98 to .99 (Diken et al. 2012). The scale has also been shown to distinguish individuals with ASD symptoms from asymptomatic groups.

Development and Well-Being Assessment (DAWBA)

Developed by Goodman and colleagues (2000), this tool has a four-component structure: interviews with children aged 5-16, parent interviews, teacher scales, and computer-based ratings. The tool includes structured and open-ended questions, allowing for the simultaneous interpretation of information. This diagnostic tool's sensitivity was found to be 92% and its specificity was found to be 89%. It was adapted into Turkish by Dursun (2009). Each section of this form consists of four questions. Structured interviews are advantageous for large groups because they are easy to administer and less costly (Dursun 2009).

Autism Behavior Checklist (ABC)

The Autism Behavior Checklist (ABC), developed by Krug and colleagues (1980) for the assessment of ASD, has been updated twice (Krug et al. 1993, 2008). The 2008 version of the ABC can be used for screening, diagnosis, and educational planning (Özdemir 2014). The 1993 scale consists of five subscales and 57 items: sensory, relating, body and object use, language skills, and social and self-help skills (Özdemir 2014). Children who scored 68 or higher on a scale ranging from 0 to 159 were considered to have a "high probability of ASD" (Eaves and Williams 2006, Özdemir 2014). Marteleto and Pedromônico (2005) reported that the scale accurately diagnosed individuals with ASD 81.6% of the time. The Turkish validity and reliability studies of the 1993 version of the scale, titled "Autism Behavior Checklist," were conducted by Yılmaz-Irmak et al. (2007), and the reliability coefficient was calculated as .86 among teachers and .59 among parents and teachers. It was found that 53 of the 57 items in the scale could distinguish between ASD and intellectual disability, and the cutoff score for the Turkish version of the scale was calculated as 39, unlike the original. With a cutoff score of 39, the sensitivity value was calculated as .82 and the specificity value as .74. Krug et al. (2008), on the other hand, excluded 10 items from the original scale from the 2008 version and included the remaining 47 items as is. Özdemir (2014) conducted the validity and reliability studies of the 2008 scale version and found a correlation of .99 for teachers and .76 for parents. The reliability correlation coefficients between the raters, consisting of teachers and parents, were calculated as .55.

Direct Observation and Self-Assessment Tools

Childhood Autism Rating Scale (CARS)

The first scale, developed by Schopler and Reichler (1971), was named the Childhood Psychosis Rating Scale. It was later revised and renamed the Childhood Autism Rating Scale (Schopler et al. 1980, Schopler et al. 2010, İncekaş-Gassaloğlu et al. 2016). The 15-item scale addresses topics such as children's relating to people, imitation, emotional response, listening response, body use, object use, adaptation to change, visual response, listening response, taste, smell, touch response, fear or nervousness, verbal communication, and nonverbal communication. It is also a tool used for both screening and diagnostic purposes, determining the severity of autism as mild-moderate or moderate-severe (Hergüner and

Özbaran 2010, Schoppler et al. 2010, İncekaş-Gassaloğlu et al. 2016). Additionally, it was developed to distinguish between children with intellectual disabilities who do not have autism and those who exhibit autistic symptoms (Marcus and Schoppler 2007, Hergüner and Özbaran 2010, İncekaş-Gassaloğlu et al. 2016). Studies indicate that the scale is also effective in distinguishing children diagnosed with autism from those who are trainable (İncekaş-Gassaloğlu et al. 2016, Kılınç et al. 2019). Bishop et al. (2008) stated that CARS is more effective for children aged 2-5. Each item on the scale is scored on a scale of 1-4. Children who score between 15-29.5 on the scale do not show signs of autism; children who score between 30-36.5 are classified as having mild-moderate autism, and children who score between 37-60 are classified as having severe autism (Garfin et al. 1988, Mesibov et al. 1989, Volkmar and Wiesner 2009, İncekaş-Gassaloğlu et al. 2016). While the adaptation of the scale to Turkish was carried out by Sucuoğlu et al. (1996), validity and reliability studies were conducted by İncekaş-Gassaloğlu et al. (2016) in a larger sample, and Cronbach's α coefficient was determined as .95, and test-retest reliability ($r = .97$, $p<.01$).

Autism Spectrum Quotient (ASQ)

The Autism Spectrum Quotient was developed by Baron-Cohen et al. (2006) to determine the extent to which an adult with average intelligence exhibits ASD characteristics. Reliability coefficients for the scale's dimensions were calculated as .82 for communication, .9 for social skills, .81 for imagination, .66 for attention to detail, and .76 for attention shifting. Reliability studies of the scale revealed adequate Cronbach's α coefficients for each dimension (Baron-Cohen et al. 2006). Adapted into Turkish by Köse and colleagues (2010), the scale is a 50-item self-report measure developed to measure subthreshold ASD characteristics in typically developing individuals. It assesses five domains: social skills, attention shifting, attention to detail, communication, and imagination. The Cronbach's alpha internal consistency coefficient for the Turkish form of the scale was calculated as .64, and the test-retest reliability was .72. Furthermore, the correlation between all subscales and item scores was significant ($p<0.001$). Principal component factor analysis revealed a three-factor structure: communication/mind reading, details, and social skills (Köse et al. 2010).

Considerations in Selecting Assessment Tools

In the screening and diagnostic processes examining ASD, several essential considerations exist in selecting, applying, and evaluating measurement tools. To implement a functional screening/diagnostic process, practitioners must consider the validity and reliability of measurement tools and the developmental characteristics of the target group to whom the test will be administered. This section details the best practice recommendations for selecting assessment tools used in ASD.

Validity

Validity is the degree to which a measurement tool can measure the structure it intends to measure (Turgut 1997, Doğan 2003, Güler 2014, Toland 2021, Cohen et al. 2022, Karabay 2022,). In other words, the capacity of the measurement tool to measure ASD symptoms is a matter of validity. For example, a measurement tool with good validity can measure ASD symptoms and also distinguish between individuals on different parts of the autism spectrum. In this regard, it is of great importance to carry out this process with measurement tools that have been validated. Validity is addressed in various ways, including content validity, criterion-related validity, and construct validity (AERA 2022). It is essential for professionals using ASD tests to focus on the following points when evaluating the validity of the tests to conduct more qualified measurement and evaluation processes.

When examining validity, it is crucial to evaluate content validity first. Content validity refers to whether the test covers the symptoms of ASD (AERA 2022, Cohen et al. 2022). In other words, content validity for ASD measurement tools refers to evaluating whether the test comprehensively addresses the symptoms and behavioural indicators of ASD. For a test to have content validity, it must be appropriate for the observed individual's age, developmental level, and cultural background (Howell 2017). At the same time, it must be able to accurately assess the individual's social skills, communication skills, and repetitive behaviors, in context.

The next step is to examine construct validity before administering the test. For ASD measurement tools, construct validity refers to how well the test reflects the theoretical structure and dimensions of ASD (Erkuş 2012). For example, whether the theoretical model established in the measurement tool contributes to ASD as a general factor is indicated in the construct validity section. Experts should carefully examine whether the test accurately reflects the relationships between different ASD symptoms and characteristics. Construct validity is determined using statistical methods such as factor analysis (Erkuş 2012). During the application process, mainly when self-report assessment scales are used, researchers/practitioners can contribute to the quality of the application process by checking whether the items work together before any hypothesis analysis.

Finally, criterion validity is another point that experts evaluating individuals with ASD should consider. Criterion-related validity assesses how well test results correspond to the actual conditions of individuals diagnosed with ASD. In this evaluation process, the test results are considered about other measurement tools that are accepted as reference/gold standards (Erkuş 2012). For example, suppose the results of a test show a high correlation with other accepted diagnostic tools such as the M-CHAT. In that case, this provides evidence supporting the criterion-related validity of the test (Lord et al. 1999, Robins et al. 2001, Zumbo and Hubley 2021).

Reliability

Reliability is the ability of a measurement tool to produce consistent results over time and under different conditions. High reliability means similar results are obtained when the same tool is used on the same individuals at other times or by various administrators. Reliability can be assessed in different ways, including test-retest reliability, internal consistency, and split-half reliability (AERA 2022). Experts using ASD-related tests should consider the following reliability characteristics.

First, test-retest reliability provides essential evidence of the reliability of a measurement tool. Test-retest reliability shows whether the test produces consistent results when administered at different times (Güler 2014, Cohen et al. 2022, Doğan 2023). Experts can evaluate the reliability of the test by comparing the results of repeated test administrations after a specific time interval. Test-retest reliability is typically reported using the Pearson correlation coefficient (r) or the intraclass correlation coefficient (ICC). A reliability of .70 or higher is generally considered acceptable, while a range of .80 to .90 is deemed sufficient for psychometric tests (Nunnally 1978). However, the correlation is expected to decrease as the time interval increases. Long intervals may include individual changes, so test-retest reliability may not solely depend on measurement error (Anastasi and Urbina 1997). Therefore, the appropriate time interval must be determined according to the administered test. For example, in test-retest applications conducted on the Autism Diagnostic Observation Schedule II (ADOS-2), it was found that the test results were highly consistent after a specific time interval. This provides strong evidence that the ADOS-2 can consistently assess ASD symptoms over time (Lord et al. 2012).

Secondly, another desired characteristic of a measurement tool is internal consistency. Experts should prefer tests that demonstrate high internal consistency. Internal consistency refers to how consistent the test items are with the general concept being tested (Kutlu 2010, AERA 2022, Karabay 2022). Statistical methods such as Cronbach's alpha coefficient or McDonald's Omega are frequently used to assess a test's internal consistency (Antalyalı and Alparslan 2020). It is recommended that Cronbach's alpha coefficient and McDonald's Omega value be at least .70 (Nunnally 1978). However, in some cases, such as newly developed scales, the .60–.70 range is also acceptable (Hair et al. 2014). For example, Robins et al. (2014) reported that the internal consistency value of the fourth revision of the M-CHAT, which they used to assess ASD, was above .90. This result proves that the test reliably measures social communication and interaction skills related to ASD.

Thirdly, experts can examine the split-half reliability coefficient of the tests. This reliability method checks whether one half of the test yields consistent results with the other half. In the test, a high correlation between the split halves is expected. Generally, a correlation of .70 or higher is acceptable (Kline 1999). Split-half reliability is considered a good indicator of the overall reliability of the test (AERA 2022). For

example, Krug et al.(1993) calculated the split-half reliability coefficient for the ABC as .87 using the Pearson Correlation method and .94 using the Spearman-Brown method in their study on the reliability of the ABC. Volkmar and colleagues (1988) determined the split-half reliability of the total score for the same scale to be .74. Researchers have demonstrated the reliability of scales using this method.

Another critical point to consider when selecting or conducting tests in the screening stages of ASD is the findings related to inter-rater reliability. This method examines whether evaluators obtain similar results when they apply the same test (Turgut 1997, Güler 2014, Baykul 2015). In assessing conditions such as ASD, the fact that different experts reach similar results can be considered an essential indicator of the reliability of the test. For example, the inter-rater reliability correlation coefficient of the CARS, which has a Cronbach's alpha coefficient of .95, was found to be .98. The results mentioned above provide evidence that the scale can be applied consistently by different professionals and produce reliable results.

Test Suitability

Test suitability refers to the suitability of the test for the group to which it will be administered; in other words, the suitability of the measurement tool for the target population (Erkuş 2012). The developmental characteristics of the target age group must be considered when designing ASD diagnostic and screening tools. For example, the CARS was developed to assess ASD symptoms from early childhood onwards, and validity and reliability studies have been conducted in Turkish (Sucuoğlu et al. 1996). The first criterion to consider is the age range of the test selected by the expert and the developmental characteristics of the target population. Another critical factor is that ASD diagnostic and screening tests should be appropriate for the language and cultural characteristics of the population in which they are administered (Yiğitoğlu and Odluyurt 2021). In addition, the tests used in ASD diagnostic and screening processes must be adapted to individuals' specific needs and circumstances. For example, the ABC, which provides the opportunity to assess individuals' communication and social interaction characteristics comprehensively, stands out as a measurement tool that meets this requirement by evaluating individuals' behavioral characteristics. It is essential to carefully assess each of these criteria to identify individuals with ASD and plan appropriate interventions correctly. This meticulousness in selecting measurement tools will increase the accuracy and effectiveness of diagnosis and screening processes.

Measurement Errors, Sources, and Possible Precautions in the Implementation of Assessment Tools

As with any measurement tool, errors can be observed while administering ASD diagnostic and screening tests. It is essential for practitioners to be aware of these measurement errors and to take precautions against them before, during, and after the assessment. This section covers the most common types of measurement errors and their sources. It also provides practitioners with recommendations for best practices to address these errors.

Random error, a type of error caused by unexpected and uncontrollable factors during the measurement process, is one of the types of errors that may be encountered during the administration of ASD tests (Pekin Çetin and Güler 2018, AERA 2022, Çıkrıkçı 2024). Random errors, which are inevitable due to the nature of the measurement process and can arise from various causes, can affect the repeatability of the measurement. For example, factors such as the individual's mood, health status, attention level, and fatigue on the day of the test are among the factors that cause random errors. The individual's performance on the test and the reliability and validity of the results can be negatively affected by each factor. To reduce random errors, these factors must be minimized (Bademci 2008). Measures such as test planning, pre-test preparation, and consideration of the individual's circumstances help to limit such errors (Anastasi and Urbina 1997, AERA 2014). In addition, it is essential to assess the physical and psychological condition of the individual being tested and to determine the most appropriate time for measurement (Erkuş 2012).

Another necessary measurement error in administering ASD tests arises from the test administrator. Such errors can significantly affect the assessment process's accuracy, objectivity, and standardization. Errors from the test administrator may be due to the administrator's lack of training, inexperience, or subjective judgments (Çelik 2021). In particular, situations such as the administrator misinterpreting the individual's responses, incorrectly applying the test instructions, or evaluating the individual's behavior based on their own biases are examples of such errors (Börkan et al. 2017). Johnson et al. (2017) noted that the education level and experience of practitioners significantly affect test results' accuracy in children diagnosed with ASD, emphasizing that practitioner education and experience are critical factors in the reliability and accuracy of test results. Suppose a test administrator incorporates their personal opinions or value judgments into the responses given by the individual during the test administration. In that case, this can negatively affect the assessment's objectivity and reduce the results' reliability. In addition, using language that is not appropriate for the individual's level of comprehension during the test or failing to provide sufficient explanation can negatively affect test results, or the administrator may act based on preconceptions related to factors such as the individual's socioeconomic status, ethnic origin, or gender, leading to test results that do not accurately reflect the individual's actual condition (Şimşek 2012). Providing test administrators with comprehensive pre-training is the most effective way to reduce errors caused by the administrator. This training should focus on correctly applying test instructions, objectively evaluating the individual's responses, and avoiding including subjective judgments in the evaluation. On the other hand, consultation, experience sharing, and supervision among experts also play a crucial role. Experienced practitioners mentoring less experienced colleagues and providing them with feedback can help reduce errors. Regular meetings or working groups where practitioners can discuss the challenges encountered during test administration will also be helpful for more qualified measurement and evaluation processes (Sayın 2010).

Another type of error in measurement is the physical conditions of the test environment (Börkan et al. 2017). The environment in which the test is administered should enable the individual to maintain their attention, feel comfortable, and respond naturally. Conditions such as noisy environments, inadequate lighting, or uncomfortable temperatures can negatively affect participants' concentration and, consequently, their test performance (Meyer et al. 2001). Especially in tests that require sensitive evaluations, such as ASD, the quality of the test environment can play a decisive role in the accuracy of the results. To prevent such errors, the test environment should be designed to minimize distractions, be quiet, and be arranged in a way that is appropriate for the individual's sensory sensitivities, thereby facilitating focus on the assessment process. Bringing physical conditions to a level where the individual can respond naturally helps prevent measurement errors while increasing the reliability and accuracy of test results (Bellini and Akullian 2007).

Another type of error that can hinder the screening and diagnosis processes of ASD is errors arising from materials and tools. First, the tests used to assess individuals accurately must be sufficient in terms of validity and reliability (DeVellis and Thorpe 2021). Culturally and linguistically inappropriate tests may produce results that do not reflect the individual's abilities. Materials that are not appropriate for the individual's age, developmental level, and sensory sensitivities can also distract the individual or cause stress and inhibit natural responses, thereby reducing the reliability of the measurement process. In addition, the physical form of the test materials is also essential. Test administrators must thoroughly examine the materials and tools before the test and have all the necessary materials ready during the test (Brockly 2013). For example, worn, incomplete, or non-standard test materials can affect the individual's performance during the test and lead to measurement errors (Çelik 2021). Therefore, test materials and tools should be checked regularly, and damaged or missing materials should be replaced or repaired promptly (Meyer et al. 2001). Control questions in measurement tools are a method used to measure consistency by presenting similar or contradictory statements multiple times in different forms. At the same time, control questions are also used to determine whether the participant has filled out the scale randomly or systematically incorrectly (Ergin 1995, Gökdemir and Yılmaz 2023). Control questions increase the validity and reliability of the data obtained from the scale (Podsakoff et al. 2003).

Another critical point is standardization. Standardization ensures that a test is administered to each individual under the same conditions and that the results are evaluated consistently. Standardization errors in ASD screening and diagnosis processes can prevent the test from producing consistent results across different groups or over time (Filipek et al. 2000). For example, if a test is developed for individuals in a specific culture and is applied to individuals in another culture without being adapted, it may lead to misleading results (Oosterhuis et al. 2017). Therefore, diversity should be considered in test development and adaptation processes, and attention should be paid to how individuals from different cultural and linguistic groups may perceive and respond to the test (Zumbo and Hubley 2021). Individuals' socioeconomic status should also be considered (Geisinger and McCormick 2013). Furthermore, unclear application protocols or protocols that administrators can interpret differently can also lead to standardization errors (Karabay 2022). For example, failure to adhere to time limits or presenting instructions in different ways can affect individuals' performance and weaken the validity of the results. To prevent standardization errors, tests must be adapted to the target group for which they will be administered and subjected to validity and reliability analyses.

Another important consideration in studies on ASD assessments is that reporting should be transparent and methodologically sound (Aksoy and Çakmak 2022). Transparent and methodologically sound reporting contributes to a better understanding of the results. To achieve transparent and methodologically sound reporting, the measurement tools used, sample characteristics, data collection, and analysis processes should be reported in detail (APA 2020). The limitations encountered in the study should be clearly stated, and the possible effects of these limitations on the results should be discussed (Mitchell and Maxwell 2013). Finally, the study results should be supported by practical recommendations and contribute to the literature. For example, guidance can be provided to educators and practitioners on improving testing processes.

Conclusion

This review comprehensively examines the measurement and evaluation tools used to screen and diagnose ASD, addressing their validity, reliability, and measurement errors. The tools examined were categorized as level 1 screening, level 2 screening, and diagnostic scales, and each measurement tool was evaluated for its psychometric properties. Furthermore, key considerations for selecting these tools and precautions for potential measurement errors during implementation are discussed.

Assessment of ASD begins with level 1 screening. Level 1 screening tools are used to screen children with early symptoms for developmental disabilities or risk of ASD in large populations. These tools are generally administered by teachers, parents, or healthcare professionals. These scales include the Denver Developmental Screening Test-II, Ankara Developmental Screening Inventory, Gazi Early Childhood Assessment Tool, Ages and Stages Questionnaire, and the Three-Item Direct Observation Screen, all of which have been validated in our country. These screening tools, also used in our country, provide preliminary screening for individuals at developmental risk, particularly in early childhood. They apply to large samples with their high internal consistency and cultural adaptability.

The evaluation of children exhibiting symptoms of ASD or identified as at risk for ASD is referred to as level 2 screening. The tools used in level 2 screening allow for the provision of multi-source data by combining parental information with observational findings. Level 2 screening transforms the "suspicion" from the first step into a clinically meaningful risk assessment. In this respect, level 2 tools are crucial in transitioning to the diagnostic phase. Tools for which validity and reliability studies have been conducted in our country include CHAT, M-CHAT, Autism Spectrum Screening Questionnaire, Social Communication Questionnaire, Social Responsiveness Scale-2, and the Communication and Symbolic Behavior Scales. Another essential function of tools at this stage is to produce data consistent with clinical diagnostic tools. For example, suppose a child identified as high-risk on the M-CHAT screening also exhibits similar severity of symptoms on the CARS or GARS-2 diagnostic process. In that case, this supports the validity of the diagnostic process. Therefore, level 2 tools serve not only a screening function but also a preliminary validation function that increases the effectiveness of diagnostic tools.

Diagnostic tools confirm the diagnosis and rank symptom severity during the evaluation of individuals identified as at risk for ASD during level 1 and level 2 screening. These tools are the final step in the diagnostic process and serve as the primary determinant in structuring individualized intervention plans. In this context, diagnostic tools enhance the validity of the diagnosis by integrating information from screening tools with objective, standardized, and psychometrically sound criteria. The tools used in our country include the Gilliam Autism Rating Scale-II, the Developmental and Well-Being Assessment, the Autism Behavior Checklist, the Autism Spectrum Quotient, and the Childhood Autism Rating Scale. However, adaptation, validity, and reliability studies for tools such as the ADOS-2 and ADI-R, which are considered "gold standards" in the international literature and play a critical role in clinical validation processes, are still limited. This hinders the ability to support the diagnostic process with universal criteria. Therefore, comprehensive validity and reliability studies of these tools are needed in Türkiye. In particular, the structured social interaction and play scenarios of the ADOS-2 hold great potential for directly confirming symptom patterns identified in level 2 screening. The ADI-R, on the other hand, offers a practical framework for revealing the developmental timeline of ASD symptoms by systematically inquiring about family history.

Measurement errors, such as practitioners' lack of training, inadequate materials, inappropriate testing environments, and random variables, can affect the validity and reliability of assessment tools used in the screening and diagnosis. Therefore, when selecting tools, it is recommended to examine pre- and post-implementation validity and reliability results thoroughly, select assessment tools appropriate for age and developmental level, ensure cultural validity, standardize practitioner training, and provide supervision support when necessary, tailor the measurement environment to the child's sensory sensitivities, and utilize multiple measurement sources that can provide data from various sources, such as parents, teachers, and observations. In summary, the selection of assessment tools used in the screening and diagnosis processes of ASD should be sensitive to technical validity and reliability criteria and the cultural, environmental, and assessment resource-related conditions within the application context. This will strengthen early diagnosis opportunities, increase the accuracy of intervention programs, and maximize children's developmental opportunities.

References

AERA (2014). Standards for Educational and Psychological Testing. American Washington DC, Educational Research Association.

AERA (2022) Standards for Educational and Psychological Testing, updated ed. Washington DC, American Educational Research Association.

Aile ve Sosyal Politikalar Bakanlığı (2023) Otizm Spektrum Bozukluğu Olan Bireylere Yönelik II. Ulusal Eylem Planı 2023-2030. Ankara, Aile ve Sosyal Politikalar Bakanlığı.

Aksoy F, Çakmak Z (2022) OSB değerlendirme: Değerlendirme sonuçlarını yorumlama ve rapor yazma. In Otizm Spektrum Bozukluğunda Değerlendirme (Ed V Aksoy):395-414. Ankara, Vize Akademik.

Aksoy V, Şahin Ş (2017). Otizm spektrum bozukluğu (OSB): Tarama, tanılama ve değerlendirme. In Zihin Yetersizliği ve Otizm Spektrum Bozukluğu (Ed İH Diken, H Bakkaloğlu):230-259. Ankara, Pegem Akademi.

APA (1994) Diagnostic and Statistic Manual Mental Disorders, 4th ed. (DSM-IV). Washington DC, American Psychiatric Association.

APA (2013) Diagnostic and Statistic Manual Mental Disorders, 5th ed. Washington DC, American Psychiatric Association.

APA. (2020). Publication Manual of the American Psychological Association, 7th ed. Washington DC, American Psychological Association.

Anastasi A, Urbina S (1997) Psychological Testing, 7th ed. Upper Saddle River (NJ), Prentice Hall.

Antalyalı ÖL, Alparslan AM (2020) Jamovi Uygulamalı İstatistik Analizleri Sosyal Bilimcilere Yönelik Araştırma Hikayeleri ile Anlatım. Ankara, Nobel Akademik.

Avcıoğlu H (2007) Sosyal becerileri değerlendirme ölçeğinin geçerlik ve güvenirlik çalışması (4-6 yaş). Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, 7:93-103.

Avcıoğlu H, Sazak-Pınar E, Öztürk T (2005) Kaynaştırma uygulamaları: Okul öncesi eğitim kurumlarında uygulanan kaynaştırmaya yönelik öğretmen ve anne-baba tutumlarının incelenmesi. In XIV. Ulusal Özel Eğitim Kongresi Bildiri Kitabı (Ed A Yıkmış, ES Pınar):69-89. Ankara, Kök Yayınevi.

Avcil S, Baykara B, Baydur H, Münir KM, Emiroğlu Nİ (2015) 4-18 yaş aralığındaki otistik bireylerde sosyal iletişim ölçeği Türkçe formunun geçerlik ve güvenilirlik çalışması. Turk Psikiyatri Derg, 26:56-64.

Bademci V (2008). Araştırmalarda ölçme ile ilgili bazı büyük hataları düzeltmek ve eğitimde yeniden yapılanmayı sürdürmek: güvenirlik, testlerin bir özelliği değildir. Endüstriyel Sanatlar Eğitim Fakültesi Dergisi, 22:50-69

Bakkaloğlu H, Kudret ZB, Ökcün-Akçamuş MÇ, Yalçın S, Demir S (2021) Turkish adaptation of Social Responsiveness Scale-2-Preschool Turkish Form: a reliability and validity study. Çukurova Üniversitesi Eğitim Fakültesi Dergisi, 50:535-560.

Baron-Cohen S, Allen J, Gillberg C (1992) Can autism be detected at 18 months? The needle, the haystack, and the CHAT. Br J Psychiatry, 161:839-843.

Baron-Cohen S, Hoekstra RA, Knickmeyer R, Wheelwright S (2006) The Autism-Spectrum Quotient (AQ)—Adolescent version. J Autism Dev Disord, 36:343-350.

Baron-Cohen S, Wheelwright S, Cox A, Baird G, Charman T, Swettenham J et al. (2000) Early identification of autism by the Checklist for Autism in Toddlers (CHAT). J R Soc Med, 93:521-525.

Baykan S, Temel ZF, Ersoy Ö, Avcı N (2002) Erken Çocukluk Gelişimi Değerlendirme Aracının GEÇDA geliştirilme süreci. In Erken Çocukluk Gelişimi ve Eğitimi Sempozyumu-Geleceğe Bakış (Ed OB Salcı):70-73. Ankara, Kök Yayınevi.

Baykul Y (2015) Eğitimde ve Psikolojide Ölçme: Klâsik Test Teorisi ve Uygulaması. Ankara, Pegem Akademi.

Bellini S, Akullian J (2007) A meta-analysis of video modeling and video self-modeling interventions for children and adolescents with autism spectrum disorders. Except Child, 73:264-287.

Bishop DV, Whitehouse AJ, Watt HJ, Line EA (2008) Autism and diagnostic substitution: evidence from a study of adults with a history of developmental language disorder. Dev Med Child Neurol, 50:341-345.

Börkan B, Öztemür G, Yılmaz O, Çetintaş Ş, Gülcan B, Özcan M (2017) Rehberlik ve araştırma merkezlerinde psikolojik test uygulama süreçleri. Türk Psikolojik Danışma ve Rehberlik Dergisi, 7:161-185.

Bricker D, Squires J, Kaminski R, Mounts L (1988) The validity, reliability, and cost of a parent-completed questionnaire system to evaluate at-risk infants. J Pediatr Psychol, 13:55-68.

Bricker D, Squires J, Mounts L, Potter L, Nickel R, Twombly E, Farrell J (1999) Ages and Stages Questionnaire. Baltimore MD, Paul H. Brookes.

Brockly ME (2013) The role of test administrator and error (Master's thesis). USA, Purdue University.

Berument SK, Rutter M, Lord C, Pickles A, Bailey A (1999) Autism screening questionnaire: diagnostic validity. Br J Psychiatry, 175:444-451.

Centers for Disease Control and Prevention (2025) Data & statistics on autism spectrum disorder. https://www.cdc.gov/mmwr/volumes/74/ss/ss7402a1.htm (Accessed 02.06.2025).

Charman T, Baird G (2002) Practitioner review: Diagnosis of autism spectrum disorder in 2-and 3-year-old children. J Child Psychol Psychiatry, 43:289-305.

Cohen RJ, Swerdlik ME, Sturman ED (2022) Psychological Testing and Assessment: An Introduction to Tests and Measurement, 10th ed. New York, McGraw-Hill.

Constantino JN (2021) Social responsiveness scale. In Encyclopedia of Autism Spectrum Disorders, (Ed FR Volkmar):4457-4467. Cham, Springer.

Constantino JN, Gruber CP (2012) Social Responsiveness Scale, 2nd ed. (SRS-2). Torrance CA, Western Psychological Services.

Corsello CM (2005) Early intervention in autism. Infants Young Child, 18:74-85.

Çakıcı A (2020) Otizm spektrum bozukluğu olan çocukların kaynaştırma/bütünleştirme yoluyla eğitim sorunlarına genel bir bakış. Boğaziçi Üniversitesi Eğitim Dergisi, 37:81-106.

Çelik C (2021) Türkiye'de WISC-IV uygulayıcı eğitimleri ve kursiyer uygulama hatalarının gözden geçirilmesi: WISC-IV Uygulayıcı Hataları. Klinik Psikoloji Dergisi, 5:183-201.

Çıkrıkçı RN (Ed) (2024) Eğitimde Ölçme ve Değerlendirme. Ankara, Pegem Akademi.

Dawson G, Rogers S, Munson J, Smith M, Winter J, Greenson J et al. (2010) Randomized, controlled trial of an intervention for toddlers with autism: the Early Start Denver Model. Pediatrics, 125:e17-e23.

DeVellis RF, Thorpe CT (2021) Scale Development: Theory and Applications. Thousand Oaks, CA, Sage.

Diken IH, Bayhan P, Turan F, Sipal RF, Sucuoglu B, Ceber-Bakkaloglu H et al. (2012) Early childhood intervention and early childhood special education in Turkey within the scope of the developmental system approach. Infants Young Child, 25:346-353.

Doğan N (2003) Eğitimde Ölçme ve Değerlendirme. Ankara, Pegem Akademi.

Doğan N, Başokçu TO, Bilgen ÖB, Yakar L, Uyar Ş, Gündeğer C et al. (2023) Eğitimde Ölçme ve Değerlendirme. Ankara, Anı Yayıncılık.

Dumont-Mathieu T, Fein D (2005) Screening for autism in young children: The Modified Checklist for Autism in Toddlers (M-CHAT) and other measures. Ment Retard Dev Disabil Res Rev, 11:253-262.

Dursun OB (2009) Gelişim ve Ruhsal Sağlık Değerlendirmesi Türkçe'ye uyarlama ve geçerlik çalışması (Doktora tezi). İzmir, Dokuz Eylül Üniversitesi.

Eaves RC, Milner B (1993) The criterion-related validity of the Childhood Autism Rating Scale and the Autism Behavior Checklist. J Abnorm Child Psychol, 21:481-491.

Eaves RC, Williams TO (2006) The reliability and construct validity of ratings for the Autism Behavior Checklist. Psychol Sch, 43:129-142.

Ehlers S, Gillberg C, Wing L (1999) A screening questionnaire for Asperger syndrome and other high-functioning autism spectrum disorders in school-age children. J Autism Dev Disord, 29:129-141.

Elder JH, Brasher S, Alexander B (2016) Identifying the barriers to early diagnosis and treatment in underserved individuals with autism spectrum disorders (ASD) and their families: a qualitative study. Issues Ment Health Nurs, 37:412-420.

Ercan İ, Kan İ (2006) Ölçme ve ölçmede hata. Eskişehir Teknik Üniversitesi Bilim ve Teknoloji Dergisi A-Uygulamalı Bilimler ve Mühendislik, 7:51-56.

Ergin DY (1995) 1. ölçeklerde geçerlik ve güvenirlik. Marmara Üniversitesi Atatürk Eğitim Fakültesi Eğitim Bilimleri Dergisi, 7:125-148.

Erkuş A (2012) Psikolojide Ölçme ve Ölçek Geliştirme. Ankara, Pegem Akademi Yayınları.

Filipek PA, Accardo PJ, Ashwal S, Baranek GT, Cook Jr EH, Dawson G et al. (2000) Practice parameter: screening and diagnosis of autism: report of the Quality Standards Subcommittee of the American Academy of Neurology and the Child Neurology Society. Neurology, 55:468-479.

Frankenburg WK, Dodds J, Archer P; Brescnik B, Maschka P, Edelman N et al. (1990) Denver II: Screening Manual. Denver, CO, Denver Developmental Materials.

Frankenburg WK, Dodds J, Archer P, Shapiro H, Bresnick B (1992) The Denver II: a major revision and restandardization of the Denver Developmental Screening Test. Pediatrics, 89:91-97.

Frankenburg WK, Dodds JB (1967) The Denver Developmental Screening Test. J Pediatr, 71:181-191.

French L, Kennedy EM (2018) Annual research review: early intervention for infants and young children with, or at-risk of, autism spectrum disorder: a systematic review. J Child Psychol Psychiatry, 59:444-456.

Garfin DG, McCallon D (1988) Validity and reliability of the Childhood Autism Rating Scale with autistic adolescents. J Autism Dev Disord, 18:367-378.

Geisinger KF, McCormick C (2013) Testing and assessment in cross-cultural psychology. In Handbook of Psychology: Assessment Psychology, 2nd ed. (Eds JR Graham, JA Naglieri, IB Weiner):114-139. Hoboken, NJ, Wiley.

Gilliam JE (1995) GARS: Gilliam Autism Rating Scale. Austin, TX, Pro-Ed.

Goodman R, Ford T, Richards H, Gatward R, Meltzer H (2000) The Development and Well-Being Assessment: description and initial validation of an integrated assessment of child and adolescent psychopathology. J Child Psychol Psychiatry, 41:645-655.

Gökdemir F, Yılmaz T (2023) Likert tipi ölçekleri kullanma, modifiye etme, uyarlama ve geliştirme süreçleri. Journal of Nursology, 26:148-160.

Güler S (2014) Eğitimde Ölçme ve Değerlendirme. Ankara, Nobel Yayınları.

Günden UO, Diken İH (2022) Otizm spektrum bozukluğu: tarama, tanılama ve gelişimsel değerlendirme. In Erken Çocuklukta Otizm Spektrum Bozukluğu (Ed İH Diken):165-201. Ankara, Pegem Akademi.

Hair JF, Black WC, Babin BJ, Anderson RE (2014) Multivariate Data Analysis, 7th ed. Upper Saddle River, Pearson.

Hensch TK (2016) The power of the infant brain. Sci Am, 314:64-69.

Hergüner S, Özbaran B (2010) Çocuk ve Ergen Psikiyatrisinde Ölçütler ve Ölçekler: Yaygın Gelişimsel Bozukluklar, Yıkıcı Davranım Bozuklukları. Ankara, Nobel Yayınevi

Howell DC (2017) Psikoloji için İstatistiksel Metotlar (Çeviri Ed. Y Baykul). Ankara, Pegem Akademi.

Huerta M, Bishop SL, Duncan A, Hus V, Lord C (2012) Application of DSM-5 criteria for autism spectrum disorder to three samples of children with DSM-IV diagnoses of pervasive developmental disorders. Am J Psychiatry, 169:1056-1064.

Hyman SL, Levy SE, Myers SM, Council on Children with Disabilities, Section on Developmental and Behavioral Pediatrics (2020) Identification, evaluation, and management of children with autism spectrum disorder. Pediatrics, 145:e20193447.

Işıl O, Kemaloğlu Y, Gökdoğan Ç, Gündüz B, Bilgin C (2015) İşitme kayıplı çocukların gelişimsel alanlardaki performans düzeylerinin Gazi Erken Çocukluk Değerlendirme Aracı ile incelenmesi. Gazi Üniversitesi Gazi Eğitim Fakültesi Dergisi, 34:563-582.

İncekaş-Gassaloğlu S, Baykara B, Avcil S, Demiral Y (2016) Çocukluk Otizmi Derecelendirme Ölçeği Türkçe Formunun geçerlik ve güvenilirlik çalışması. Turk Psikiyatri Derg, 27:266-274.

Johnson S, Hollis C, Kochhar P, Hennessy E, Wolke D, Marlow N (2010) Autism spectrum disorders in extremely preterm children. J Pediatr, 164:525-531.

Kapçı EG, Küçüker S, Uslu S (2010) How applicable is the Ages and Stages Questionnaires to Turkish children? Topics in Early Childhood Special Education, 30:176-188.

Kara B, Mukaddes NM, Altınkaya I, Güntepe D, Gökçay G, Özmen M (2014) Using the Modified Checklist for Autism in Toddlers in a well-child clinic in Turkey: adapting the screening method based on culture and setting. Autism, 18:331-338.

Karabay E (2022) Korelasyon-ölçme araçlarında bulunması gereken psikometrik özellikler. In Eğitimde Ölçme ve Değerlendirme (Eds B Gök, Ö Bıkmaz-Bilgen):15-42. İstanbul, Lisans Yayıncılık.

Kılınç Ç, Bağlama B, Akçamete G (2019) Otizm spektrum bozukluğunun erken çocukluk döneminde tanılanma ve değerlendirilmesinde kullanılan ölçme araçlarının incelenmesi. Kıbrıs Türk Psikiyatri ve Psikoloji Dergisi, 1:200-205.

Kline P (1999) The Handbook of Psychological Testing, 2nd ed. London, Routledge.

Korkmaz E (2009) 6-24 aylardaki normal gelişim gösteren ve gelişimsel gecikmesi olan bebeklerin iletişim ve sembolik davranışlarının karşılaştırılması (Yüksek lisans tezi). Ankara, Hacettepe Üniversitesi.

Köse S, Bora E, Erermiş S, Aydın C (2010) Otizm-Spektrum Anketi Türkçe formunun psikometrik özellikleri. Anadolu Psikiyatri Derg, 11:253-260.

Köse S, Özbaran B, Yazgan Y, Bildik T, Erermiş S, Aydın C et al. (2017) 6-18 yaş aralığındaki çocuklarda Otizm Spektrum Tarama Ölçeği'nin Türkçe uyarlamasının psikometrik özellikleri. Turk Psikiyatri Derg, 28:268-277.

Krug DA, Arick J, Almond P (1980) Behavior checklist for identifying severely handicapped individuals with high levels of autistic behavior. J Child Psychol Psychiatry, 21:221-229.

Krug DA, Arick J, Almond P (1993) Autism Screening Instrument for Educational Planning. Austin, TX, Pro-Ed.

Krug DA, Arick JR, Almond PJ (2008) Autism Screening Instrument for Educational Planning, Examiner's Manual. Austin, TX, Pro-Ed.

Kurnaz-Adıbatmaz FB, Özyürek A (2019) Erken çocukluk döneminde gelişimsel değerlendirmenin önemi: GEÇDA sonuçları örneği. Bingöl Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 9:1321-1343.

Kutlu Ö, Doğan CD, Karakaya İ (2010) Eğitimde Ölçme ve Değerlendirme: Performansa ve Portfolyoya Dayalı Durum Belirleme. Ankara, Pegem Akademi.

Landa RJ (2008) Diagnosis of autism spectrum disorders in the first 3 years of life. Nat Clin Pract Neurol, 4:138-147.

Lord C, Rutter M, DiLavore PC, Risi S, Gotham K, Bishop S (2012) Autism Diagnostic Observation Schedule, Second Edition (ADOS-2). Torrance CA, Western Psychological Services.

Lord C, Rutter M, DiLavore PC, Risi S (1999) Autism Diagnostic Observation Schedule: Manual. Torrance CA, Western Psychological Services.

Marcus L, Schopler E (2007) Educational approaches for autism—TEACCH. In Clinical Manual for the Treatment of Autism (Eds E Hollander, E Anagnostou):211-233. Washington DC, American Psychiatric Publishing.

Marteleto MRF, Pedromônico MRM (2005) Validity of Autism Behavior Checklist (ABC): preliminary study. Braz J Psychiatry, 27:295-301.

Matson JL, Boisjoli JA, Hess JA, Wilkins J (2010) Factor structure and diagnostic fidelity of the Baby and Infant Screen for Children with aUtIsm Traits–Part 1 (BISCUIT–Part 1). Dev Neurorehabil, 13: 72-79.

Mesibov GB, Schopler E, Schaffer B, Michal N (1989) Use of the Childhood Autism Rating Scale with autistic adolescents and adults. J Am Acad Child Adolesc Psychiatry, 28:538-541.

Meyer GJ, Finn SE, Eyde LD, Kay GG, Moreland KL, Dies RR et al. (2001) Psychological testing and psychological assessment: a review of evidence and issues. Am Psychol, 56:128.

Mitchell MA, Maxwell SE (2013) A comparison of the cross-sectional and sequential designs when assessing longitudinal mediation. Multivariate Behav Res, 48:301-339.

Nunnally JC (1978) Psychometric Theory, 2nd ed. New York, McGraw-Hill.

Öner P, Öner Ö, Münir K (2014) Three-item Direct Observation Screen (TIDOS) for autism spectrum disorder. Autism, 18:733-742.

Oosterhuis HE, van der Ark LA, Sijtsma K (2017) Standard errors and confidence intervals of norm statistics for educational and psychological tests. Psychometrika, 82:559-588.

Özdemir O (2014) Otizm Davranış Kontrol Listesi Türkçe versiyonu geçerlik ve güvenirlik çalışmaları (Doktora tezi). Eskişehir, Anadolu Üniversitesi.

Pekin Z, Çetin S, Güler N (2018) Otizm Sosyal Beceriler Profili Ölçeğinde puanlayıcılar arası güvenirliğin farklı kuramlara göre karşılaştırılması. Journal of Measurement and Evaluation in Education and Psychology, 9:202-215.

Podsakoff PM, MacKenzie SB, Lee JY, Podsakoff NP (2003) Common method biases in behavioral research: a critical review of the literature and recommended remedies. J Appl Psychol, 88:879.

Robins DL, Casagrande K, Barton M, Chen CMA, Dumont-Mathieu T, Fein D (2014) Validation of the Modified Checklist for Autism in Toddlers, Revised with Follow-Up (M-CHAT-R/F). Pediatrics, 133:37-45.

Robins DL, Fein D, Barton ML, Green JA (2001) The Modified Checklist for Autism in Toddlers: An initial study investigating the early detection of autism and pervasive developmental disorders. J Autism Dev Disord, 31:131-144.

Rogers SJ (2005) Play interventions for young children with autism spectrum disorders. In Empirically Based Play Interventions for Children (Eds LA Reddy, TM Files-Hall, CE Schaefer):215-239. Washington DC, American Psychological Association.

Rutter M, Bailey A, Lord C (2003) The Social Communication Questionnaire: Manual. Torraqnce, CA, Western Psychological Services.

Savaşır I, Sezgin N, Erol N (1994) Handbook of Ankara Developmental Screening Inventory. Ankara, Türk Psikologlar Derneği.

Sayın S (2010) Bilimsel araştırmalarda yapılan istatistiksel ve yöntembilimsel hatalar-II: Grafik, tablo ve gösterim hataları. Türk Eğitim Bilimleri Dergisi, 8:117-143.

Schopler E, Reichler RJ (1971) Parents as cotherapists in the treatment of psychotic children. J Autism Child Schizophr, 1:87-102.

Schopler E, Reichler RJ, DeVellis RF, Daly K (1980) Childhood autism rating scale. J Autism Dev Disord, 10:91-103.

Schopler E, Van Bourgondien ME, Wellman, GJ, Love SR (2010). Childhood Autism Rating Scale, 2nd ed. Torrance, CA, Western Psychological Services.

Selimoğlu ÖG, Özdemir S, Töret G, Özkubat U (2013) Otizmli çocuğa sahip ebeveynlerin otizm tanılama sürecinde ve tanı sonrasında yaşadıkları deneyimlerine ilişkin görüşlerinin incelenmesi. International Journal of Early Childhood Special Education, 5:129-161.

Sucuoğlu B, Akkök F, Öktem F, Gökler B (1996) Otistik çocukların değerlendirilmesinde kullanılan ölçeklere ilişkin bir çalışma. 3P Psikiyatri, Psikoloji, Psikofarmakoloji Dergisi, 4:117-121.

Şimşek A (2012) Sosyal Bilimlerde Araştırma Yöntemleri. Eskişehir, Anadolu Üniversitesi.

Şişman HZ (2021) Riskli bebeklerin "Alberta Infant Motor Skalası" ve "Erken Gelişim Evreleri Envanteri" ile gelişiminin değerlendirilmesi (Yüksek lisans tezi). İstanbul, İstanbul Üniversitesi.

Temel F, Ersoy Ö, Avcı N, Turla A (2004) Gazi Erken Çocukluk Değerlendirme Aracı. Ankara, Rekmay.

Tetik-Kabil A (2005) Türk örneklemi ile yapılan Erken Çocukluk Dönemi Otizm Tarama Ölçeği geçerlik çalışması (Yüksek lisans tezi). İstanbul, Boğaziçi Üniversitesi.

Toland MD (2014) Practical guide to conducting an item response theory analysis. J Early Adolesc, 34:120-151.

Turgut F (1997) Eğitimde Ölçme ve Değerlendirme Metotları. Ankara, Saydam Matbaası.

Turhan B, Özbay Y (2016) Erken çocukluk eğitimi ve nöroplastisite. Uluslararası Erken Çocukluk Eğitimi Çalışmaları Dergisi, 1:54-63.

Uran-Kurtgöz P (2023) Otizm spektrum bozukluğu: tarama ve tanılama (Tıbbi Değerlendirme). In Tanıdan Müdahaleye Otizm Spektrum Bozukluğu El Kitabı (Eds B Sucuoğlu, H Bakkaloğlu, MC Ökcün-Akcamuş):101-120. Ankara, Vize Akademik.

Ünden UO, Diken İH (2022) Otizm spektrum bozukluğu: tarama, tanılama ve gelişimsel değerlendirme. In Erken Çocuklukta Otizm Spektrum Bozukluğu (Ed İH Diken):165-201. Ankara, Pegem Akademi.

Volkmar FR, Wiesner LA (2009) A Practical Guide to Autism: What Every Parent, Family Member, and Teacher Needs to Know. Hoboken, NJ, Wiley. .

Volkmar FR, Cicchetti DV, Dykens E, Sparrow SS, Leckman JF, Cohen DJ (1988) An evaluation of the Autism Behavior Checklist. J Autism Dev Disord, 18:81-97.

Wetherby AM, Prizant BM (2002) Communication and Symbolic Behavior Scales: Developmental Profile. Baltimore MD, Paul H. Brookes.

WHO (2024) Clinical Descriptions and Diagnostic Requirements for ICD-11 Mental, Behavioural and Neurodevelopmental Disorders. Geneva, World Health Organisation.

Yalaz K, Anlar B, Bayoğlu B (2011) Denver II Gelişimsel Tarama Testi "Türkiye Standardizasyonu". Ankara, Gelişimsel Çocuk Nörolojisi Derneği.

Yazıcı DN, Akman B (2020) The expectations of parents having children with autism from early intervention programs. International Journal of Early Childhood Special Education, 12:264-276.

Yıkgeç A (2005) A validity study of the Modified Checklist for Autism in Toddlers (M-CHAT) on a Turkish sample (Yüksek lisans tezi). İstanbul, Boğaziçi Üniversitesi.

Yılmaz-Irmak T, Tekinsav-Sütçü S, Aydın A, Sorias O (2007) Otizm Davranış Kontrol Listesinin (ABC) geçerlik ve güvenirliğinin incelenmesi. Çocuk ve Gençlik Ruh Sağlığı Dergisi, 14:13-23.

Yiğitoğlu E, Odluyurt S (2021) Otizm spektrum bozukluğu tanısı koyan doktorların tıbbi tanılama sürecine ilişkin görüşlerinin incelenmesi. Ankara Üniversitesi Eğitim Bilimleri Fakültesi Özel Eğitim Dergisi, 22:1-29.

Zumbo BD, Hubley AM (2021) Understanding and Investigating Response Processes in Validation Research. New York, Springer.

Zwaigenbaum L, Bauman ML, Stone WL, Yirmiya N, Estes A, Hansen RL et al. (2015) Early identification of autism spectrum disorder: Recommendations for practice and research. Pediatrics, 136(Suppl 1):10-40.

Authors Contributions: The author(s) have declared that they have made a significant scientific contribution to the study and have assisted in the preparation or revision of the manuscript
Peer-review: Externally peer-reviewed.
Ethical Approval: This review study does not require ethical clearance.
Conflict of Interest: No conflict of interest was declared.
Financial Disclosure: No financial support was declared for this study..