Reliability of The Childhood Autism Rating Scale, Second Edition Assessments Utilizing Telepsychiatry for Children with Neurodevelopmental Disorders and Their Caregivers: Randomized Feasibility Study

Shunya Kurokawa, Kensuke Nomura, Nana Hosogane, Takashi Nagasawa, Yuko Kawade, Yu Matsumoto, Shuichi Morinaga, Yuriko Kaise, Ayana Higuchi, Akiko Goto, Naoko Inada, Masaki Kodaira, Taishiro Kishimoto

Submitted to: Journal of Medical Internet Research
on: August 21, 2025

Table of Contents

Reliability of The Childhood Autism Rating Scale, Second Edition Assessments Utilizing Telepsychiatry for Children with Neurodevelopmental Disorders and Their Caregivers: Randomized Feasibility Study

Shunya Kurokawa[1] MD, PhD; Kensuke Nomura[2, 3] MD, PhD; Nana Hosogane[4] MD; Takashi Nagasawa[5] MD; Yuko Kawade[2, 6] MD; Yu Matsumoto[2] MD; Shuichi Morinaga[7] MD; Yuriko Kaise[1] MA; Ayana Higuchi[1] MA; Akiko Goto[6] MD, PhD; Naoko Inada[8] MA, PhD; Masaki Kodaira[4]; Taishiro Kishimoto[1] MD, PhD

[1] Center for the Promotion of Interdisciplinary Research in Medicine and Life Science, Keio University School of Medicine Tokyo JP
[2] Department of Child Psychiatry, Shimada Ryoiku Medical Center for Challenged Children Tokyo JP
[3] Department of Neuropsychiatry, School of Medicine, Keio University Tokyo JP
[4] Department of Child and Adolescent Mental Health, Aiiku Clinic Tokyo JP
[5] Tokyo Metropolitan Children's Medical Center Tokyo JP
[6] Tsurugaoka Garden Hospital Tokyo JP
[7] Department of Child and Adolescent Psychiatry, Tokyo Metropolitan Children's Medical Center Tokyo JP
[8] Department of Clinical Psychology, Taisho University Tokyo JP

Corresponding Author:
Taishiro Kishimoto MD, PhD

Center for the Promotion of Interdisciplinary Research in Medicine and Life Science, Keio University School of Medicine
Mori JP Tower F7, 1-3-1, Azabudai, Minato-ku
Tokyo
JP

Abstract

Background: The coronavirus disease (COVID-19) pandemic has increased the demand for telepsychiatry in neurodevelopmental assessment; however, few studies have examined the inter-rater reliability of clinician-rated autism scales, such as the Childhood Autism Rating Scale, Second Edition (CARS2), when used remotely. Although the feasibility and validity of telehealth assessments have been explored previously, the consistency between remote and in-person CARS2 ratings by different clinicians remains underinvestigated.

Objective: To evaluate the inter-rater reliability of the CARS2 scores administered via telepsychiatry compared with in-person assessments in children with autism spectrum disorder (ASD) and/or attention-deficit/hyperactivity disorder (ADHD).

Methods: In this randomized feasibility study, 74 children (mean age 10.4 years) with DSM-5 diagnoses of ASD, ADHD, or both were recruited from five medical institutions in Japan. Each participant underwent two CARS2 evaluations, one face-to-face and the other via telepsychiatry, by two independent evaluators who were blinded to each other's ratings. Remote assessments were conducted using a secure video platform facilitated by the caregiver. Inter-rater reliability was assessed using intraclass correlation coefficients (ICC) stratified by diagnosis, sex, and age.

Results: The overall ICC for CARS2 total scores between face-to-face and remote assessments was 0.582 (95% CI 0.407–0.715), indicating moderate agreement. The ICCs were slightly higher in children with ASD (0.541) than in those with ADHD (0.495). Reliability was greater among girls (ICC 0.666) than boys (0.459), and among older children (?11 years, ICC 0.676) compared to younger children (?10 years, ICC 0.417). Qualitative observations and prior literature suggest that these differences may reflect age- and context-related behavioral variability, challenges in observing non-verbal communication via video, and environmental distractions in home settings.

Conclusions: Remote administration of the CARS2 via telepsychiatry yielded moderate overall inter-rater reliability with better consistency in older children and girls. Although remote tools show promise for increasing access to developmental assessments, clinicians should remain aware of context- and modality-specific limitations. Further work with larger and more diverse samples

should refine the protocols and enhance standardization. Clinical Trial: The study protocol was registered with the UMIN Clinical Trials Registry (UMIN000039860).

(JMIR Preprints 21/08/2025:82784)
DOI: https://doi.org/10.2196/preprints.82784

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http
No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <a href="https:/

Original Manuscript

Original Paper

Reliability of The Childhood Autism Rating Scale, Second Edition Assessments Utilizing Telepsychiatry for Children with Neurodevelopmental Disorders and Their Caregivers: Randomized Feasibility Study

Shunya Kurokawa, MD, PhD;[1] Kensuke Nomura, MD, PhD;[1,2] Nana Hosogane, MD;[3] Takashi Nagasawa, MD;[4] Yuko Kawade, MD;[2,5] Yu Matsumoto, MD;[2] Shuichi Morinaga, MD;[1,6] Yuriko Kaise, MA;[1] Ayana Higuchi, MA;[1] Akiko Goto, MD, PhD;[5] Naoko Inada, MA, PhD;[7] Masaki Kodaira, MD, PhD;[3] Taishiro Kishimoto, MD, PhD[1]*

Corresponding author: Taishiro Kishimoto (tkishimoto@keio.jp)

[1]Center for Promotion of Interdisciplinary Research in Medicine and life Science, Keio University School of Medicine, Tokyo, Japan
[2]Department of Child Psychiatry, Shimada Ryoiku Medical Center for Challenged Children, Tokyo, Japan
[3]Department of Child and Adolescent Mental Health, Aiiku Clinic, Tokyo, Japan
[4]Department of Child and Adolescent Psychiatry, Tokyo Metropolitan Children's Medical Center, Tokyo, Japan
[5]Tsurugaoka Garden Hospital, Tokyo, Japan
[6]Hiratsuka City Hospital, Kanagawa, Japan
[7]Department of Clinical Psychology, Taisho University, Tokyo, Japan

Abstract

Background: The coronavirus disease (COVID-19) pandemic has increased the demand for telepsychiatry in neurodevelopmental assessment; however, few studies have examined the inter-rater reliability of clinician-rated autism scales, such as the Childhood Autism Rating Scale, Second Edition (CARS2), when used remotely. Although the feasibility and validity of telehealth assessments have been explored previously, the consistency between remote and in-person CARS2 ratings by different clinicians remains underinvestigated.
Objective: To evaluate the inter-rater reliability of the CARS2 scores administered via telepsychiatry compared with in-person assessments in children with autism spectrum disorder (ASD) and/or attention-deficit/hyperactivity disorder (ADHD).
Methods: In this randomized feasibility study, 74 children (mean age 10.4 years) with DSM-5 diagnoses of ASD, ADHD, or both were recruited from five medical institutions in Japan. Each participant underwent two CARS2 evaluations, one face-to-face and the other via telepsychiatry, by two independent evaluators who were blinded to each other's ratings. Remote assessments were conducted using a secure video platform facilitated by the caregiver. Inter-rater reliability was assessed using intraclass correlation coefficients (ICC) stratified by diagnosis, sex, and age.
Results: The overall ICC for CARS2 total scores between face-to-face and remote assessments was 0.582 (95% CI 0.407–0.715), indicating moderate agreement. The ICCs were slightly higher in children with ASD (0.541) than in those with ADHD (0.495). Reliability was greater among girls (ICC 0.666) than boys (0.459), and among older children (≥11 years, ICC 0.676)

compared to younger children (≤10 years, ICC 0.417). Qualitative observations and prior literature suggest that these differences may reflect age- and context-related behavioral variability, challenges in observing non-verbal communication via video, and environmental distractions in home settings.

Conclusions: Remote administration of the CARS2 via telepsychiatry yielded moderate overall inter-rater reliability with better consistency in older children and girls. Although remote tools show promise for increasing access to developmental assessments, clinicians should remain aware of context- and modality-specific limitations. Further work with larger and more diverse samples should refine the protocols and enhance standardization.

Trial Registration: The study protocol was registered with the UMIN Clinical Trials Registry (UMIN000039860).

Keywords: attention-deficit hyperactivity disorder; autism spectrum disorder; telepsychitatry

Introduction

Autism spectrum disorder (ASD) is a neurodevelopmental condition characterized by persistent deficits in social communication and interaction alongside restricted, repetitive patterns of behavior [1]. The global prevalence of ASD is estimated to be approximately 1–2% of the population, though reported rates vary by region and have generally increased in recent decades [2]. Accurate diagnosis and assessment of ASD severity typically requires a combination of direct behavioral observations of the child and in-depth caregiver interviews.

The coronavirus disease (COVID-19) pandemic has accelerated the adoption of telepsychiatry for ASD evaluation, prompting extensive research on the reliability and validity of remote diagnostic tools. Early studies and reviews have generally been encouraging; telepyschiatric autism evaluations show a high agreement with in-person diagnoses and have been well received by families and clinicians [3,4]. These findings suggest that when appropriately adapted, remote ASD assessment methods can feasibly bridge gaps in access without compromising diagnostic accuracy.

Structured caregiver interviews, such as the ADI-R or the Diagnostic Interview for Social and Communication Disorders are particularly amenable to remote use because they rely on verbal reports rather than live child observations. Because these interviews relied on parent-reported histories rather than on tasks administered directly to the child, their core content could be preserved over the video.

For example, telepsychiatric administration of the ADI-R has shown strong agreement with in-person evaluations. An U.S. clinical trial of 51 toddlers (18–36 months) found that a caregiver-mediated tele-assessment achieved 86% agreement with standard face-to-face ADI-R ASD diagnoses [5]. Notably, a randomized study likewise reported no differences in diagnostic consistency between in-person vs. video-based ADI-R assessments; inter-rater reliability on ADI-R scores was statistically equivalent across modes [6]. These findings indicate that the ADI-R diagnostic algorithm remains robust when interviews are conducted using a video.

In contrast, adapting direct behavioral observation tools such as the Autism Diagnostic Observation Schedule (ADOS) for telehealth has been more challenging. The ADOS is a semi-structured play-based assessment often regarded as a “gold standard” for ASD diagnosis, but it traditionally requires face-to-face interaction with the child. When clinicians attempted to administer the ADOS over

video, they could achieve approximately 86% concordance in overall diagnostic decisions compared with in-person evaluations. However, item-level reliability was markedly lower for remote behavioral observations: in one study, the caregiver interview portion (ADI-R) maintained substantial inter-rater agreement (κ ~0.82) via telehealth, whereas the remote ADOS items showed only moderate agreement (κ ~0.50) [6].

However, certain telehealth adaptations of the ADOS have yielded promising results under specific conditions. For instance, a modified remote administration of the ADOS for verbally fluent adults demonstrated an intraclass correlation of approximately 0.92 for diagnostic classification, comparable to an in-person evaluation [7].

These mixed outcomes illustrate that, while remote behavioral observations are feasible, ensuring high reliability for all components of an autism assessment battery can be difficult when conducted online.

Beyond full diagnostic interviews and observational batteries, more streamlined ASD rating scales have been explored for remote assessment. The Childhood Autism Rating Scale, Second Edition (CARS2) is one such tool that has attracted interest in the telehealth context . The CARS2 is a clinician-rated scale that quantifies ASD symptom severity across 15 domains and is valued for its brief administration time and minimal training requirements [8]. Traditionally, the CARS2 scoring integrates information from multiple sources (caregiver reports, clinical records, and direct observation); however, it can also be based on focused direct observation of the child alone.

During the pandemic, this instrument was adapted for remote use by having clinicians observe and interact with children via videoconference, an approach often termed the remote CARS2 (rCARS2) assessment. A recent validation study of rCARS2 against in-person ADOS-2 reported promising results: remote CARS2 scores were strongly correlated with ADOS-2 scores (Spearman ρ ~0.64) [9]. These findings support the utility of CARS2 as an efficient telehealth-compatible observation tool for ASD, capable of maintaining good concordance with gold-standard in-person assessments.

Despite a growing body of literature on telehealth-based ASD assessments, a notable gap remains in our understanding of the interrater reliability of these remote assessment tools. To date, most tele-ASD studies have focused on validity (i.e., agreement with in-person diagnoses or standard instruments) and feasibility, whereas far fewer have examined whether different clinicians observing the same child, either in person or via video, assign consistent ratings. Interrater reliability is a critical property of any diagnostic instrument that reflects consistency and objectivity across examiners. Notably, we are unaware of any prior work that has specifically evaluated the inter-rater reliability of the CARS2 when administered via telepsychiatry.

However, prior research suggests that clinician-rated scales of attention-deficit/hyperactivity disorder (ADHD) can be implemented reliably via telehealth: with substantial agreement between telepsychiatry and face-to-face assessments [10]. Building on this foundation, the present study aimed to evaluate the inter-rater reliability of the CARS2 administered via telepsychiatry versus in-person. Clarifying the reliability of remote CARS2 assessments is clinically important for establishing trusted telehealth practices in ASD diagnosis and ensuring that evaluations yield consistent results regardless of whether they occur face-to-face or online.

Methods

Participants

Patients were recruited from Keio University Hospital and four collaborating institutions (Shimada Ryoiku Medical Center for Challenged Children, Aiiku Clinic, Tokyo Metropolitan Children's Medical Center, and Tsurugaoka Garden Hospital). The inclusion criteria were as follows: (1) confirmed Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) diagnoses of ASD; (2) age between 6 and 17 years at the time of obtaining consent; and (3) stable pharmacotherapy, if applicable, for at least 3 months before consent.

Exclusion criteria included: (1) hearing or visual impairment in either the child or caregiver that interfered with evaluation despite corrective devices; (2) absence of caregivers knowledgeable about the participants' early developmental history; (3) presence of severe comorbid psychiatric symptoms (e.g., hallucinations and delusions) as determined by clinicians; and (4) initiation of new treatments, such as pharmacotherapy or psychotherapy, anticipated during the study period. The study protocol was approved by the Ethics Committee of Keio University School of Medicine (20190301). This study was registered in the UMIN Clinical Trial Registry (UMIN000039860).

Baseline Assessments

Demographic and clinical data, including age, sex, diagnosis, medication history, illness duration, and intelligence test results, were extracted from medical records. Baseline assessments utilized the Autism-Spectrum Quotient–Japanese version for children, the Strengths and Difficulties Questionnaire to evaluate social adjustment, and the Short Sensory Profile to assess sensory characteristics.

Study Procedure

The CARS2 was used as the primary measure to evaluate the inter-rater reliability between different evaluators in face-to-face assessments. The CARS2 includes two versions: the Standard Version (CARS2-ST) and the High-Functioning Version (CARS2-HF), with the latter recommended for children aged ≥6 years who have an IQ >80 or relatively intact verbal abilities. In designing this study, we assumed patients without comorbid intellectual disability.

Following the baseline assessments, each participant underwent two independent CARS2 evaluations conducted face-to-face by two different evaluators. To minimize recall bias and practice effects, the order of the evaluators was randomized and each evaluator was blinded to the other's ratings. The interval between the two evaluations ranged from a minimum of two weeks to a maximum of three months.

The evaluators were licensed psychologists trained in developmental assessment. Prior to the study, evaluators underwent standardized training and demonstrated adequate inter-rater reliability in five pilot cases, confirming the consistency and accuracy of the CARS2 tool.

In this study, we set structured prompts to help examiners avoid missing key behaviors and to minimize the effects of different raters. To minimize modality-specific bias while sampling a broad range of social communication behaviors, we embedded three structured observation blocks—UNO play, semi-structured social interviews, and reverse interviews using photo prompts— along with the standard CARS2-QPC.

UNO play with a caregiver elicited spontaneous turn-taking, joint attention, gesture use, and affect, allowing direct scoring of CARS2 Items: 1 Relating to People, 2 Imitations, 3 Emotional Responses, 7 Visual Responses, 12 Non-verbal Communication, and 13 Activity Levels.

The semi-structured social interview followed a fixed sequence of anchor questions (hobbies, chores, school strengths/weaknesses, family pets, sources of happiness, sadness, fear, anger, and loneliness) with body-state probes, such as “How does your body feel when you are happy?. These standardized topic shifts left room for natural dialogue, providing comparable material for Item 3 Emotional Response, Item 6 Adaptation to Change, Item 10 Fear or Nervousness, Item 11 Verbal Communication, Item 12 Non-verbal Communication, and Item 14 Level/Consistency of Intellectual Response.

Next, the reverse interview asked the child to “interview the examiner” for two minutes using staged photographs (birthday party; borrowing an umbrella) as optional cues. This segment sampled spontaneous question generation, perspective-taking, and imaginative language, informing items 1, 11, 12, and 14, and the overall item 15 General Impressions. Screen-sharing ensured identical visual stimuli in the telepsychiatric condition.

The remote smartphone assessment tool “Curon®” by MICIN Co. Ltd. was used for the telemedicine evaluation. The participants were seated in front of a smartphone in their homes and introduced to a remote evaluator. Remote evaluators administered their evaluations in a room at the Keio University Hospital using a personal computer. Although it was assumed that a download and upload environment of 50Mbps or higher would be ideal for remote video applications, stable communication was achieved at approximately 5 Mbps, which in most cases fulfilled the Japanese standard 4G network. Assessments began after confirming that there were no interruptions in the video or audio environment.

Statistical Analysis

Inter-rater reliability was assessed using intraclass correlation coefficients (ICC) for continuous CARS2 scores, categorized according to standard performance parameters: almost perfect (0.81–1.00); substantial (0.61–0.80); moderate (0.41–0.60); fair (0.21–0.40); and slight (0.0–0.20). A sample size of at least 31 participants was calculated to achieve an 80% power, assuming an ICC of 0.8 and aiming for a 95% CI width of 0.3. Normally distributed data are reported as mean (SD), and categorical variables as numbers (%). Statistical analyses were performed using SPSS software (version 25.0).

Ethical Considerations

The ethics committee of Keio University School of Medicine approved the study protocol (approval number 20190301). This study was registered in the UMIN Clinical Trial Registry (UMIN000039860). Written informed consent was obtained from the caregivers, and age-appropriate informed assent was obtained from the participating children. Data confidentiality was maintained through anonymization and secure storage was ensured at all times.

Results

Participants

The demographic characteristics of the participants are presented in Table 1. Of the 75 patients who consented to participate in the study, one dropped out because of time limitations. One participant with moderate intellectual disability expressed interest in participating and completed the assessment using the CARS2-ST. However, because this case was the only case using CARS2-ST, the data were excluded from the final analysis. Of the 73 participants, 2 were from Keio University Hospital, 43 from Shimada Ryoiku Medical Center for Challenged Children, 13 from Aiiku Clinic, 10 from Tokyo Metropolitan Children's Medical Center, and 6 from Tsurugaoka Garden Hospital. Of the 73 participants, 17 (23%, 17/73) were women and 56 (77%, 56/73) were men. All patients were Asians (Japanese). Their ages ranged from 6 to 16 years, with an average of 10.4 (standard deviation [SD] = 2.5). The primary clinical diagnoses were as follows: 30 individuals (41%, 30/73) were diagnosed with ASD, 43 individuals (59%, 43/73) were diagnosed with ADHD, and 22 (30%, 22/73) with comorbid ASD and ADHD. Forty-seven caregivers (77%, 47/61) had prior experience with remote video calls, while 14 (23%, 14/61) did not.

Table 1. Demographic characteristics.

	All (*n*=73)	ADHD (*n*=43)	ASD (*n*=30)
Sex (female, %)	17 (15.7%)	12 (27.9%)	5 (16.1%)
Age (years)	10.4 (2.5)	10.5 (2.4)	10.2 (2.6)
SDQ	22.06(5.25)	21.09 (5.35)	23.24 (4.96)
AQ	24.56 (7.37)	22.86 (6.72)	27.03 (7.68)
Conners 3	112.70 (38.49)	108.34 (39.94)	118.41 (36.40)
SSP	70.90 (20.37)	67.44 (20.38)	75.63 (19.73)
CARS2 (face-to-face)	35.88 (4.88)	34.19 (4.49)	38.25 (4.45)
ADHD-RS-IV total (face-to-face)	27.14 (9.91)	26.88 (10.38)	27.52 (9.35)
ADHD-RS-IV hyperactivity/impulsiveness (face-to-face)	7.99 (6.07)	7.95 (6.60)	8.03 (5.33)
ADHD-RS-IV total inattention (face-to-face)	19.15 (5.32)	18.93 (5.28)	19.48 (5.46)

Data represented as mean (SD). ADHD, attention-deficit hyperactivity disorder; ASD, autism spectrum disorder; ADHD-RS-IV, Attention-Deficit/Hyperactivity Disorder Rating Scale IV; AQ, autism spectrum quotient; SSP, Short Sensory Profile; SDQ, Strength and Difficulties Questionnaire; CARS2, Childhood Autism Rating Scale, Second Edition; SD, standard deviation

ICCs of CARS2 Between Face-to-Face and Remote Assessments

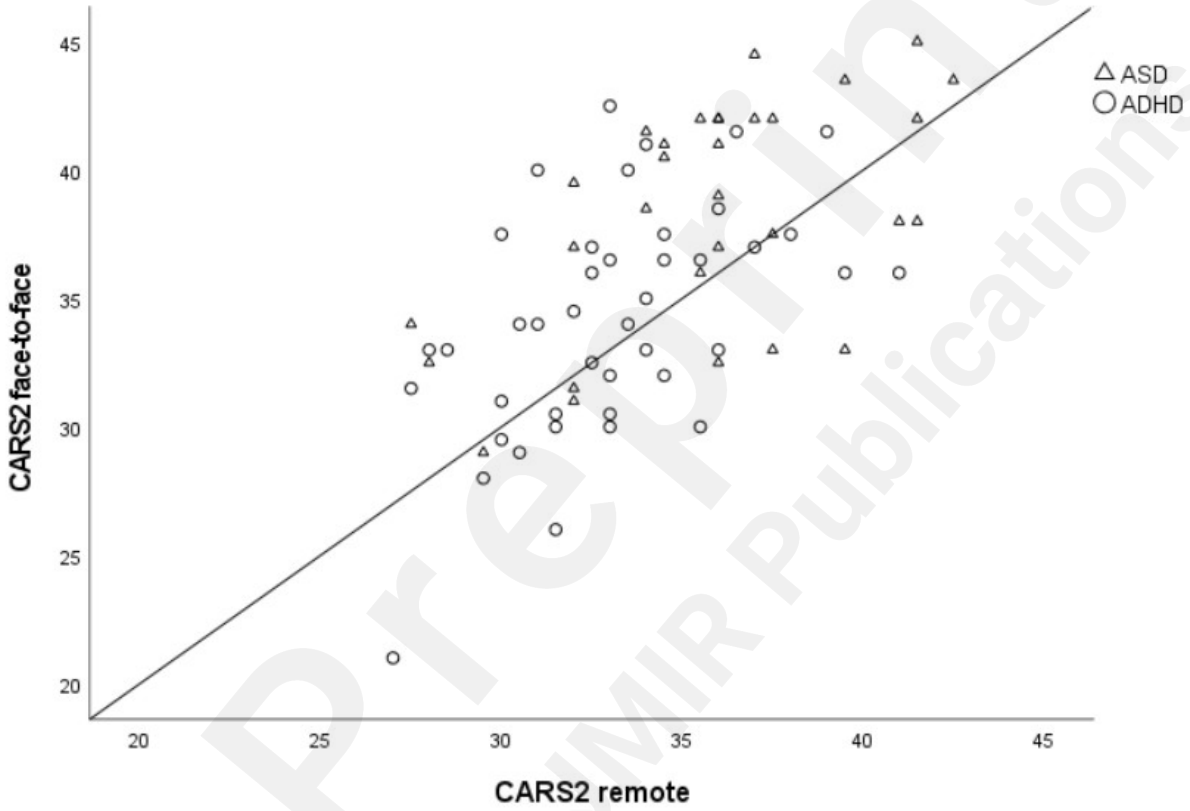
The ICC of CARS2 obtained after preliminary training was 0.885 (95% CI 0.418–0.987; *P*=.005).

The ICCs of the CARS2 between the face-to-face and remote assessments are shown in Figure 1.

The ICC of the total CARS2 was 0.582 (95% CI 0.407–0.717; *P*<.001), which was “moderate” according to the Landis criteria [11]. In patients with ADHD as their primary diagnosis, the ICC for total score was 0.495 (95% CI 0.228–0.693; *P*<.001), indicating “moderate” agreement. In patients with ASD as their primary diagnosis, the ICC for total score was 0.546 (95% CI 0.238–0.755; *P*<.001), indicating “moderate” agreement.

With regard to sex, the ICC of CARS2 in male patients was 0.495 (95% CI 0.228–0.693; *P*<.001), while the ICC was 0.666 (95% CI 0.287–0.864; *P*=.001) in the female group.

In regards to age, comparing the older age group (11 years and older, *n*=37) and the younger age group (10 years and younger, *n*=36), The ICC for total score was 0.676 (95% CI 0.433–0.828; *P*<.001) in the older age group and 0.419 (95% CI 0.127–0.644; *P*=0.003) in the younger age group.



Discussion

Principal Results

In this study, we examined the reliability of administering the CARS2 remotely via telepsychiatry compared with the traditional in-person approach. The primary finding was that the agreement between remote and face-to-face CARS2 evaluations was moderate (total score ICC ≈ 0.58). This level of concordance indicates that remote assessments generally track the same severity patterns observed in the clinic while providing a clinically acceptable degree of precision.

The agreement across subgroups followed a similar pattern. Children with ASD showed a solid moderate correspondence (ICC ≈ 0.58), whereas children with primary ADHD displayed a slightly lower—but still moderate—agreement (ICC ≈ 0.50). Girls exhibited higher reliability (ICC ≈ 0.67) than boys (ICC ≈ 0.50), and older children (≥ 11 years; ICC ≈ 0.68) outperformed younger children (≤ 10 years; ICC ≈ 0.42). These trends suggest that age and sex can influence how closely

telepsychiatric ratings mirror clinical observations.

Critically, no subgroups demonstrated poor agreement, underscoring the practical utility of remote CARS2 administration when in-person visits are difficult. Although tele-assessment is not yet a wholesale replacement for clinic-based evaluations, consistently moderate ICCs provide a positive signal that remote CARS2 scoring can serve as a dependable adjunct, especially for older youth with ASD, while clinicians remain mindful of the slightly wider variability seen in younger children and boys.

Limitations

Some limitations of our study should be acknowledged when interpreting the findings. First, our sample was modest and drawn exclusively from Japanese tertiary clinics treating children already diagnosed with ASD or ADHD, which may limit the generalizability to undiagnosed groups or different cultural settings. Also as the present study was designed around participants eligible for the CARS2-HF, the findings may not generalize to populations for whom the CARS2-ST would be more appropriate—such as children under 6 years of age or those with significant communication impairments.

Second, constraints arise from CARS2 itself. As a clinician-rated observational scale, it integrates caregiver reports, records, and direct observation, and this semi-structured flexibility invites variation: one examiner may emphasize caregiver input, while another may emphasize what is seen in the session. In the telehealth arm, parents also acted as camera operators and facilitators, creating an interaction that differed from the clinic, where the clinician directed the child. As CARS2 lacks a fixed task set, observational opportunities can vary from session to session.

Third, our ICC represented the agreement between two raters working in two different modes; it merged inter-rater variability with modality effects. We could not fully disentangle the extent of the discrepancy stemming from the telehealth medium (camera angle, bandwidth, and home setting) versus ordinary rater subjectivity. Designs using the same clinician in both modes or dual observers watching a single session would better separate these factors, but were beyond the scope of this study.

Fourth, home environments introduce uncontrollable noise, such as living rooms or shared spaces with nearby siblings, background media, and limited camera placement. Even with advanced guidance, real-world settings cannot match the controls of a clinical observation room.

In summary, a specific sample, the flexible yet subjective nature of CARS2, conflated rater-and-setting variance, and environmental constraints all suggest caution when applying moderate ICC values to other contexts.

Comparison with Prior Work

Our findings align with and add to earlier research on telepsychiatric testing for autism. The moderate match we saw between the remote and in-person CARS2 scores was almost the same size as that reported by other teams. For instance, Reese et al. found that a video caregiver interview (ADI-R) kept very high agreement ($\kappa \approx 0.82$), but the observational ADOS part fell to only moderate ($\kappa \approx 0.50$) when done remotely [12]. A newer “brief” protocol that pairs the BOSA(Brief Observation of Symptoms of Autism) with a parent interview reached $\kappa=0.66$ [13]. In contrast, adults tested using two cameras and an operator who could zoom and pan showed excellent agreement (ICC=0.92) [7]. Together, these studies show the same pattern: talking with

parents over video works very well, but rating the fine-grained behaviors you see on screen is still harder than in person unless you use high-end equipment.

The location where the test occurs is also important. Telepsychiatry allows clinicians to observe children in their everyday spaces, which can reveal their natural behavior. However, the small apartments, background noise, and parents juggling the camera made the examinations less uniform. We found lower agreement in younger children, mirroring our earlier work with telepsychiatry ADHD-RS ratings [10]. Some children were quieter or more restrained on the video, a point that matches reports that autism traits can shift with context [14].. Our data quantify how these shifts reduce scoring consistency, especially in crowded urban Japanese homes where private spaces are scarce. We also noted that parents sitting next to their child for the entire call sometimes hesitated to discuss problem behaviors openly, which could subtly change certain CARS2 item scores.

Technology adds another layer. Eye contact, facial expressions and small hand gestures are tricky to judge through a single webcam. Two-camera systems or remote-controlled units provide a clearer view but are difficult to roll out in everyday care [7,14]. However, the effects of age and sex are poorly understood. We observed higher agreement in older children, who stayed on tasks better, and girls. One guess is that the girls in our sample had more obvious symptoms, because girls are often diagnosed only when difficulties are very clear despite “camouflaging” [15]. Future studies should test whether this finding holds true for larger sample sizes.

Finally, our choice of CARS2 is a part of the picture. During COVID-19 many clinics switched to remote CARS2 because gold-standard ADOS visits were impossible [16]. Bertollo’s dissertation showed a good match between remote CARS2 and in-person ADOS ($\rho \approx 0.64$) and 83 % diagnostic accuracy [9]. What was still missing was a clear look at how well two different clinicians agree when they rate the same child in the clinic and by telepsychiatry, which is exactly what our study provides. The remaining gaps that we found point to straightforward upgrades: clearer camera guidelines, structured parent coaching, extra-short tasks to prompt key behaviors, and possibly low-cost secondary cameras. Each step should help bring remote CARS2 ratings closer to clinical standards, while maintaining telepsychiatry.

Conclusions

Remote administration of the CARS2, Second Edition, via telepsychiatry shows promise as a practical supplement to in-clinic assessments. Across our sample, total CARS2 scores obtained at home-tracked clinic scores showed moderate agreement (strongest for older children, girls, and those with ASD), and no subgroup fell below the acceptable reliability range. Although the home environment, camera setup, and flexible scoring rules introduced extra variability, these factors did not erase the core signals of autism-related behaviors. Taken together, this evidence suggests that clinicians can use CARS2 telepsychiatry sessions to triage, monitor progress, and prioritize follow-up when travel or clinical capacity is limited, provided that they remain alert to wider score swings in younger boys and children assessed primarily for ADHD.

At the same time, our findings highlight clear targets for improvement: standardized camera guidelines, parent coaching, and simple scripted activities that elicit key behaviors could further tighten reliability. Larger and more diverse cohorts, especially undiagnosed children, and study designs that isolate the rater and modality effects will refine these estimates. Until then, CARS2 telepsychiatry should be viewed not as a replacement for gold-standard clinic evaluations but as a reliable adjunct that widens access while maintaining clinically useful precision.

Funding/Support

This study was supported in part by the Japan Science and Technology Agency Program on the Open Innovation Platform with Enterprises, Research Institute, and Academia (JST-OPERA) (grant number JPMJOP1842) and MICIN Inc., Tokyo, Japan.

Role of the Funder/Sponsor

The funding sources had no role in the design and conduct of the study; the collection, management, analysis, and interpretation of the data; the preparation, review, or approval of the manuscript; or the decision to submit the manuscript for publication.

Conflict of interest

KN received speaker honoraria from Takeda, Janssen, Eli Lilly, Eisai, MSD, Otsuka, and Shionogi. MK has received honoraria from Takeda and Shionogi. TK received consultant fees from Dainippon Sumitomo, Novartis, and Otsuka; speaker honoraria from Banyu, Eli Lilly, Dainippon Sumitomo, Janssen, MSD, Novartis, Otsuka, and Pfizer; and grant support from Takeda, Dainippon-Sumitomo, and Otsuka. The remaining authors declare no conflicts of interest.

Abbreviations

ADHD: attention-deficit hyperactivity disorder
ADHD-RS-IV: Attention-Deficit/Hyperactivity Disorder Rating Scale -IV
ADOS: Autism Diagnostic Observation Schedule
ASD: autism spectrum disorder
CARS2: Childhood Autism Rating Scale, Second Edition
CI: confidence interval
COVID-19: coronavirus disease 2019
ICC: intraclass correlation coefficients
rCARS2: remote CARS2

References

1. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. 5th edition (DSM-5). Arlington, VA: American Psychiatric Association; 2013. ISBN:9780890425541
2. Zeidan J, Fombonne E, Scorah J, et al. Global prevalence of autism: a systematic review update. Autism Res. 2022;15(5):778–790. doi:10.1002/aur.2696
3. Alfuraydan M, Croxall J, Hurt L, Kerr M, Brophy S. Use of telehealth for facilitating the diagnostic assessment of Autism Spectrum Disorder (ASD): a scoping review. PLoS One. 2020;15(7):e0236415. doi:10.1371/journal.pone.0236415
4. Stavropoulos KKM, Bolourian Y, Blacher J. A scoping review of telehealth diagnosis for autism spectrum disorder. PLoS One. 2022;17(2):e0263062. doi:10.1371/journal.pone.0263062
5. Corona LL, Weitlauf AS, Hine J, et al. Parent perceptions of caregiver-mediated

telemedicine tools for assessing autism risk in toddlers. J Autism Dev Disord. 2021;51:476–486. doi:10.1007/s10803-020-04554-9
6. Reese RM, Jamison R, Wendland M, et al. Evaluating interactive videoconferencing for assessing symptoms of autism. Telemed J E Health. 2013;19(9):671–677. doi:10.1089/tmj.2012.0312
7. Schutte JL, McCue MP, Parmanto B, et al. Usability and reliability of a remotely administered adult autism assessment, the Autism Diagnostic Observation Schedule (ADOS) Module 4. Telemed J E Health. 2015;21(3):176–184. doi:10.1089/tmj.2014.0011
8. Dawkins T, Meyer AT, Van Bourgondien ME. The relationship between the Childhood Autism Rating Scale: second edition and clinical diagnosis utilizing the DSM-IV-TR and the DSM-5. J Autism Dev Disord. 2016;46:3361–3368. doi:10.1089/tmj.2014.0011
9. Bertollo JR. Autism Assessment from Home: Evaluating the Remote Childhood Autism Rating Scale, Second Edition (rCARS2) Observation for Tele-Assessment of Autism. Dissertation. Virginia Polytechnic Institute and State University; 2024. https://vtechworks.lib.vt.edu/items/7b0c1ea7-0473-4ae5-9c5e-9055f86adca5 [accessed July 10, 2025].
10. Kurokawa S, Nomura K, Hosogane N, et al. Reliability of telepsychiatry assessments using the Attention-Deficit/Hyperactivity Disorder Rating Scale-IV for children with neurodevelopmental disorders and their caregivers: randomized feasibility study. J Med Internet Res. 2024;26:e51749. doi:10.2196/51749
11. Reese RM, Jamison TR, Pabalan DJ, et al. Evaluating interactive videoconferencing for assessing symptoms of autism in toddlers. J Autism Dev Disord. 2013;43(12):3093–3099. doi:10.1089/tmj.2012.0312
12. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977 Mar;33(1):159-74.
13. Stroupková L, Preiss M, Štěpánková H, et al. Use of telehealth in autism spectrum disorder assessment in children: evaluation of an online diagnostic protocol including the Brief Observation of Symptoms of Autism. J Autism Dev Disord. Published online September 26, 2024. doi:10.1007/s10803-024-06524-x
14. Spain D, Stewart GR, Mason D, et al. Telehealth autism diagnostic assessments with children, young people, and adults: qualitative interview study with England-wide multidisciplinary health professionals. JMIR Ment Health. 2022;9(7):e37901. doi:10.2196/37901
15. Hull L, Mandy W, Petrides KV, et al. “Putting on my best normal”: social camouflaging in adults with autism spectrum conditions. J Autism Dev Disord. 2017;47(8):2519–2534. doi:10.1007/s10803-017-3166-5
16. Micheletti M, Brukilacchio BH, Hooper-Boyle H, et al. Evaluating the efficiency and equity of autism diagnoses via telehealth during COVID-19. J Autism Dev Disord. 2025;55(5):1932–1938. doi:10.1007/s10803-023-05986-9

Supplementary Files

Figures

Intraclass correlation coefficients (ICCs) comparing telepsychiatry and in?person CARS2 total scores, shown overall and by diagnosis (ASD, ADHD). Abbreviations: CARS2, Childhood Autism Rating Scale, Second Edition; ASD, autism spectrum disorder; ADHD, attention?deficit/hyperactivity disorder.