Psychometric properties of assessment instruments for autism spectrum disorder: a systematic review of Brazilian studies

Propriedades psicométricas de instrumentos de avaliação do transtorno do espectro do autismo: uma revisão sistemática de estudos brasileiros

Bárbara Backes[1], Bruna Gomes Mônego[1], Cleonice Alves Bosa[1], Denise Ruschel Bandeira[1]

ABSTRACT

Objective: To systematically review the scientific literature on the psychometric properties of international instruments for the assessment of autism spectrum disorder (ASD) in the Brazilian population. **Methods:** A search of bibliographic references was conducted in six electronic databases: PsycINFO, PubMed, IndexPsi, Lilacs, Capes (theses and dissertations) and SciELO. The studies were selected by two independent researchers. **Results:** The procedure identified 11 studies of the Brazilian population that encompassed six ASD assessment tools. Given the information provided, the adaptation of the M-CHAT, a screening instrument, was the best conducted. All steps of the adaptation process were described and the changes made to the final version of the instrument were presented, which was not addressed in other studies. In terms of reliability, all of the instruments that assessed internal consistency showed adequate values. In addition, the ADI-R and the CARS adaptations also satisfactorily contemplated inter-rater reliability and test-retest indices, respectively. Finally, all studies aiming to validate instruments showed evidence of validity and sensitivity, and specificity values above 0.90 were observed in the ASQ, ADI-R and ABC. **Conclusion:** Considering both the psychometric aspects and the copyright information, the screening instrument that currently appears to be best indicated for clinical and research use is the M-CHAT. It was also noticed that there are still no specific ASD diagnostic tools available for use in Brazil. This lack of diagnostic instruments consists in a critical situation for the improvement of clinical practice and the development of research in this area.

RESUMO

Objetivo: Revisar sistematicamente a literatura científica acerca das propriedades psicométricas de instrumentos internacionais para a avaliação do transtorno do espectro do autismo (TEA) na população brasileira. **Métodos:** Realizou-se uma busca de referências bibliográficas em seis bases de dados: PsycINFO, PubMed, IndexPsi, Lilacs, Capes (teses e dissertações) e SciELO, sendo a seleção dos estudos realizada por dois pesquisadores independentes. **Resultados:** O procedimento resultou em 11 estudos que abarcaram seis instrumentos de ava-

Address for correspondence: Bárbara Backes
Instituto de Psicologia,
Federal University of Rio Grande do Sul
Ramiro Barcelos Street, 2600
90035-003 – Porto Alegre, RS, Brazil
Telephone/Fax: +55 51 3308-5261
E-mail: barbara.edas@gmail.com

liação do TEA na população brasileira. Dadas as informações fornecidas, a adaptação do M-CHAT, um instrumento de rastreamento, foi a melhor realizada. Todas as etapas do processo de adaptação foram descritas e as alterações feitas para a versão final do instrumento foram apresentadas, o que não foi explicitado em outros estudos. Em termos de fidedignidade, todos os instrumentos que avaliaram a consistência interna apresentaram valores adequados. A ADI-R e a CARS também contemplaram satisfatoriamente a fidedignidade entre avaliadores e os índices de teste-reteste, respectivamente. Finalmente, todos os estudos que objetivaram a validação de instrumentos apresentaram evidências de validade e sensibilidade, e valores de especificidade acima de 0,90 foram observados no ASQ, ADI-R e ABC.

Conclusão: Considerando-se os aspectos psicométricos e de direitos autorais, o instrumento de rastreamento atualmente mais adequado para uso clínico e em pesquisa é o M-CHAT. Observou-se, contudo, que ainda não há um instrumento de diagnóstico específico para o TEA disponível para uso no Brasil. Essa lacuna consiste em uma situação crítica para o aprimoramento da prática clínica e da pesquisa na área do TEA.

Palavras-chave

Transtorno autístico, avaliação, psicometria, revisão.

INTRODUCTION

Autism spectrum disorder (ASD) is a neurodevelopmental condition characterized by socio-communicative impairment and the presence of repetitive and stereotyped behavior[1]. The ASD term, present in the DSM-5, replaces the pervasive developmental disorders term, used in the DSM-IV, and encompasses autistic disorder, Asperger's syndrome, childhood disintegrative disorder and the pervasive developmental disorder not otherwise specified. All those disorders represent a single condition with three different levels of severity: (1) Requiring support; (2) Requiring substantial support; and (3) Requiring very substantial support, that should be followed by the specifiers "with or without accompanying intellectual disability", "with or without accompanying language impairment" and "associated with a known medical or genetic condition or environmental factor"[1]. Thus, the dimensional nature of the classification is stressed.

Although the etiology of the disorder has not yet been established, studies have identified genetic and neurobiological factors that tend to be associated with ASD[2,3]. Regarding epidemiology, international studies show a higher incidence in males, with a ratio of 4.2 male births for each female[4]. Prevalence is approximately one in every 88 births[5], making autism one of the most common developmental disorders[4,6]. The increasing prevalence can be explained by the expansion of the diagnostic criteria, the improvement of health services related to the disorder and the change in the age of diagnosis[4].

The diagnosis of ASD is based on a qualitative assessment of behavioral patterns and is directly influenced by the complexity and variability in the presentation of the disorder (*e.g.*, levels of severity, association with intellectual disability and other medical conditions). Such characteristics have led to the development of a significant number of international instruments focusing on identification and early diagnosis[7].

However, this number is greatly reduced in Brazil, which has led researchers to conduct psychometric studies aiming to adapt international instruments for use in Brazil. It is therefore important to consider the proper and responsible use of these instruments based on psychometric criteria and the existence of copyright.

In this sense, psychometrics, a field of measurement of psychological variables, provides numerous tools with which it is possible to investigate the suitability of instruments through validity and reliability studies[8]. However, these procedures must be preceded by the adaptation of the instrument to the environment in which it will be used and by standardizing the procedures for its use[9].

Standardization aims to ensure that the procedures involved in the administration of the instrument, including the interpretation of results, are uniform[8]. The adaptation process has a number of steps. Gjersing *et al.*[10] suggested that initially, an investigation of the conceptual equivalence of items should be conducted, followed by two independent translations of the instrument, a synthesis of these two versions, two independent back-translations and a further synthesis into a single version. This single version, in turn, should be subject to an assessment by a committee of experts and to a pre-test, then reviewed and its operational equivalence investigated. Finally, there should be a primary study and an exploratory and confirmatory analysis, from which the final instrument would originate.

Similarly, Borsa *et al.*[9] proposed an adaptation model that includes the translation of the instrument, the synthesis of the translated versions, an assessment of the synthesized version by experts, an assessment of the instrument by the target audience, a back-translation, a pilot study and an assessment of the instrument's factorial structure. For these authors, the process of adaptation is directly related to the validity and reliability of the instrument and must therefore consider cultural differences at both the conceptual and the linguistic levels.

Validity is the capacity of the instrument to properly evaluate what it intends to evaluate[8,11]. There have been discussions in this area in relation to the possible types of validity. The more traditional view, called Tripartite (content, construct and criteria validity), considers validity to be an attribute of the instrument itself[8,11]. Since 1999, however, a new vision has been widely disseminated by the American Educational Research Association, the American Psychological Association and the National Council on Measurement in Education[12]. According to this perspective, validity can be obtained through various sources in addition to the three proposed in the Tripartite view. The term "evidence of validity" is thus used to express the notion that several studies may be taken together to indicate the degree of validity of a particular instrument[8,12]. Differences aside, the primary goal is that the instruments have psychometric properties that are considered to be satisfactory for use in a complementary fashion in any assessment. Another important concept in psychometric terms is reliability, which is related to the accuracy and consistency of results and suggests how reliable an instrument is[8]. A lack of reliability implies measurement errors, defined as fluctuations in the results that are influenced by factors that are irrelevant to the assessment purpose of the instrument[8].

In addition, it is important to consider the copyright of the instruments both for clinical and research use. It is common in the field of psychometric studies that the instrument has not yet been acquired by national publishers. To conduct these studies, it is therefore necessary to obtain the permission of the publisher (or author) who manages the instrument's copyright by registering the projects; when registering, the purpose of use of the instrument and the population to be investigated must be specified, among other aspects. It should be emphasized that disrespect of copyright can lead to different outcomes and might be subject to the compensation provided by the Brazilian Copyright Law (Law nº 9,610/98). Another important aspect concerns the training demanded by some publishers, which can take place at different steps of the psychometric process, including the translation step.

As mentioned earlier, the number of ASD screening and diagnostic instruments is very limited in Brazil, representing an obstacle to the expansion of research in this field and to the quality of services. This situation has led to studies involving the translation, adaptation and validation of instruments. However, a critical examination of the psychometric quality of these studies is still lacking in Brazilian publications.

This study therefore aims to systematically review the scientific literature on the psychometric properties of international instruments for ASD assessment in the Brazilian population. More specifically, this study seeks to review the quality of the psychometric studies conducted in Brazil and to investigate the suitability of these instruments (screening, diagnosis, copyrights) for helping professionals (clinicians and researchers) in the most appropriate choice of assessment tool.

METHODS

Materials

Articles and dissertations that aimed to translate, adapt and validate international ASD assessment instruments for use in Brazil were studied. Searches were therefore conducted on national and international databases encompassing studies published until February 2014. No restrictions were applied regarding chronology or the original language of publication.

Procedures

A search of references was performed in five electronic databases: PsycINFO, PubMed, IndexPsi, Lilacs [*Literatura Latino-Americana e do Caribe em Ciências da Saúde* (Latin American and Caribbean Health Sciences Literature)], Capes [*Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (Coordination for the Improvement of Higher Level Personnel)] (theses and dissertations) and SciELO encompassing studies published until February 2014, except for the Capes database whose selected studies were published until October 2012. In the first two databases, the search had four axes, based on the terms (1) "autism" or "pervasive developmental disorder" and (2) "translating" or "validity" or "psychometrics" or "adaptation" and (3) "test" or "instrument" or "checklist" or "questionnaire" and (4) "Brazil". The search in the other four databases was performed using three axes based on the following descriptors: (1) *"autismo"* or *"transtornos globais do desenvolvimento"* and (2) *"tradução"* or *"validade"* or *"psicometria"* or *"adaptação"* and (3) *"teste"* or *"instrumento"* or "checklist" or *"questionário"* [(1) "autism" or "pervasive developmental disorders" and (2) "translation" or "validity" or "psychometry" or "adaptation" and (3) "test" or "instrument" or "checklist" or "questionnaire"]. Three or four terms were therefore used depending on the language of the database searched, cross-referencing the different axes. The search was performed cross-referencing the terms "autism" and "translation" and "test", then "autism" and "translation" and ""instrument" and so on until all possibilities were exhausted. This process resulted in 32 combinations. More than one search was performed in each database using the Boolean operator or/*ou* because these searches have lower accuracy compared to those using only the Boolean operator and/*e*.

The search results included articles, dissertations or complete theses, and they were requested directly from the corresponding authors in the cases where they could not be fully accessed. The selection of studies was based on the abstract, and data extraction was performed based on an analysis of

the full articles/dissertations. Both procedures (selection and extraction) were performed independently by two judges, co-authors of this study. In the absence of agreement as to the selected studies and extracted information, an expert was consulted to reach a consensus. Studies were excluded if (1) the sample was not Brazilian, (2) the study was not empirical, (3) it did not investigate the psychometric properties of the instrument (translation, adaptation, validation or accuracy), (4) the instrument studied did not specifically assess ASD, and (5) the instrument was not international. Figure 1 shows a detailed flowchart of the study selection process.

Analysis of information

Based on the final outcome of the selection, the studies were characterized according to their nature (paper, thesis or dissertation), instrument studied, objective, subjects and journal of publication or institution of origin. In addition, information was provided on the instruments studied, such as type of use, time and mode of administration, age group for which it is intended and copyrights statements.

RESULTS

The search for bibliographical references resulted in 350 studies being retrieved. The 11 studies that constituted the final result of this review investigated the psychometric properties of six instruments, namely the Autism Behavior Checklist (ABC), the Autism Diagnostic Interview-Revised (ADI-R), the Autistic Traits Assessment Scale (ATA), the Autism Screening Questionnaire (ASQ), the Childhood Autism Rating Scale (CARS) and the Modified Checklist for Autism in Toddlers (M-CHAT). Table 1 displays a characterization of these instruments.

Considerations on the sample used

The composition of the sample is a key feature on the development of researches[13]. It is important to notice that there was great variability between samples and data collection procedures across the reviewed studies. The number of participants ranged from 5 to 303, the collection sites included public and private institutions and collection itself was conducted in person or over the Internet. The form of administration of the instruments is also different in that ATA and CARS are based on the direct observation of the child, while the remaining instruments are administered to an adult who has contact with the child. The respondents used in the studies included parents, teachers, speech therapists and other health professionals. Table 2 displays a characterization of studies selected.

Regarding the participants' diagnostic evaluation, whether ASD, intellectual disability (ID) or another psychiatric disorder, one study reported that the diagnosis was made by one of the authors[14], another considered the diagnosis reported by the participant him/herself[15], and only one study mentioned that the diagnosis was conducted by an interdisciplinary team[16]. The other studies accessed this information from the record of contacted clinics, specialized outpatient clinics and special schools[17-19]. Only one study mentioned the intelligence quotient (IQ) of the participants with ID[14] and no study reported the IQ of the participants with ASD.
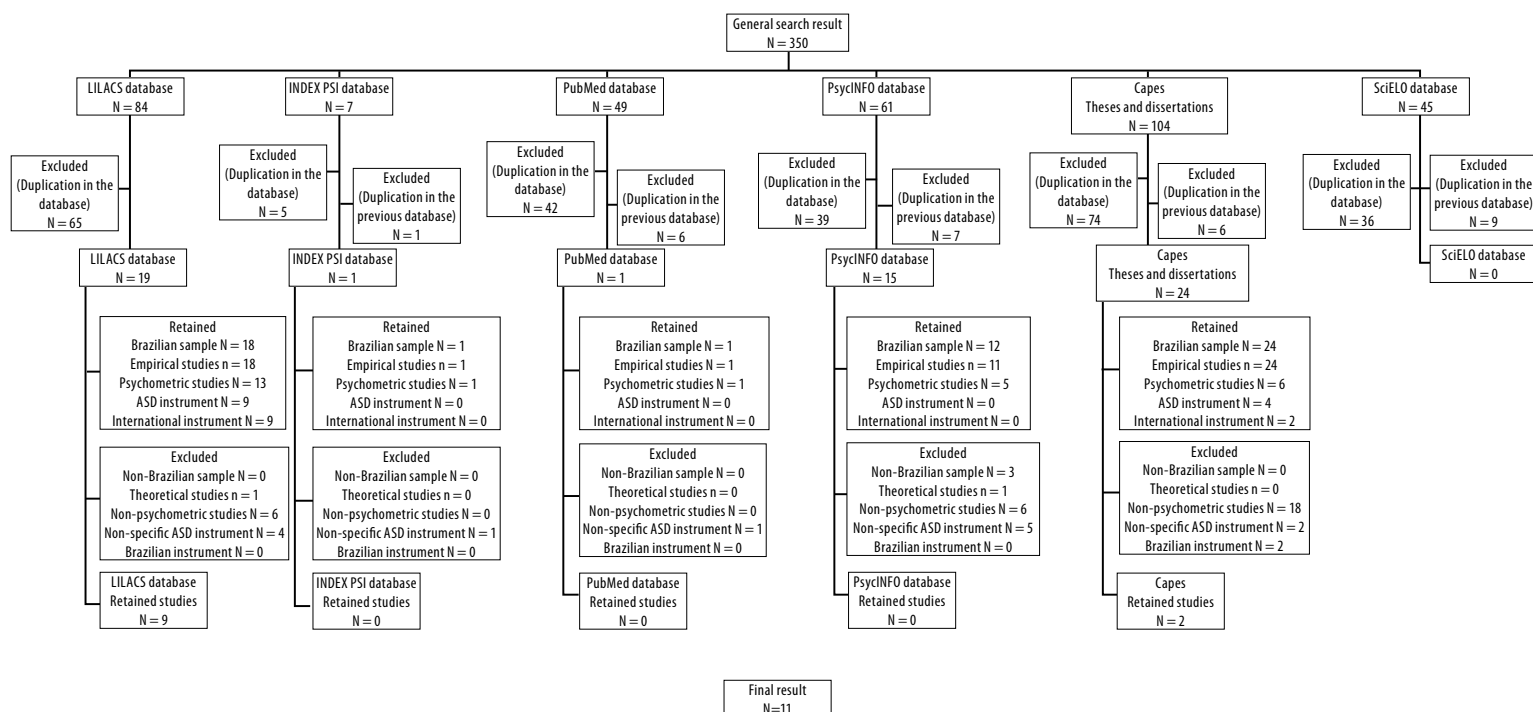

Figure 1. Flowchart of study selection.

Table 1. Characterization of the 11 studies selected

Authors (Year)	Nature of source	Instrument studied	Objective of the study	Subjects (Age)	Journal of publication/ institution	Cutoff value/sensitivity/ specificity
Aguiar (2005)*[28]	Unpublished thesis	ADI-R	Translate	5 caretakers (NR)	Universidade Presbiteriana Mackenzie	- / - / -
Assumpção Jr. et al. (1999)[17]	Article	ATA	Translate, adapt and validate	61 participants (between 2 and 19 yrs-old)	Arquivos de Neuropsiquiatria	15 / 0,96 / -
Assumpção Jr. et al. (2008)[25]	Article	ATA	Validate	93 participants (between 4 and 12 yrs-old)	Medicina de Reabilitação	23 / 0,82 / 0,75
Becker et al. (2012)[14]	Article	ADI-R	Translate and validate	40 caretakers (patients between 7 and 18 yrs-old)	Universidade Federal do Rio Grande do Sul	- / 100 / 100
Castro-Souza (2011)[15]	Unpublished thesis	M-CHAT	Adapt and validate	303 respondents (patients between 18 months and 22 yrs-old)	Universidade de Brasília	- / - / -
Losapio and Pondé (2008)[21]	Article	M-CHAT	Translate and adapt	40 respondents (NR)	Revista de Psiquiatria do Rio Grande do Sul	- / - / -
Marteleto and Pedromônico (2005)[24]	Article	ABC	Validate	133 caretakers (NR)	Revista Brasileira de Psiquiatria	68 / 57,89 / 94,73 49 / 92,6 / 92,6
Marteleto et al. (2008)[16]	Article	ABC	Validate	Group 1: 23 respondents (M = 38 yrs-old); Group 2: 15 respondents (M = 37 yrs-old)	Revista Brasileira de Psiquiatria	- / - / -
Matteo et al. (2009)[26]	Article	CARS	Validate	76 participants (between 4 and 13 yrs-old)	Medicina de Reabilitação	33 / 0,81 / 0,83
Pereira et al. (2008)[19]	Article	CARS	Translate, adapt and validate	60 respondents (patients between 3 and 17 yrs-old)	Revista Brasileira de Pediatria	- / - / -
Sato et al. (2009)[18]	Article	ASQ	Translate, adapt and validate	120 respondents: 40 children with ASD (M = 9.8 yrs-old); 40 with another psychiatric disorder (M = 11.1 yrs-old); 40 with Down syndrome (M = 9.9 yrs-old)	Revista Brasileira de Psiquiatria	14,5 / 92,5 / 95,0

NR: not reported; ABC: Autism Behavior Checklist; ADI-R: Autism Diagnostic Interview-Revised; ATA: Autistic Traits Assessment Scale; ASQ: Autism Screening Questionnaire; CARS: Childhood Autism Rating Scale; M-CHAT: Modified Checklist for Autism in Toddlers. *The content of this thesis could not be fully accessed.

Adaptation process

The adaptation process was published in detail only in the study conducted by Losapio and Pondé regarding the M-CHAT[21]. The authors stated that this process was conducted on the basis of Reichenheim and Moraes' model[22] and all of the steps were described, i.e., translation, back-translation, equivalence analysis, expert assessment and two pilot studies. This article also made clear all of the modifications made to the instrument until its final version and reported that there was a final checking by the original author of the M-CHAT.

The ADI-R was adapted for the Brazilian context by Becker et al.[14], with the permission of the copyright-owning publisher (WPS), and after completion of the training required for its use by one of the authors. The adaptation procedures were based on Sperber's model[23]. The translation was performed by two independent translators and the back-translation was sent to the original authors of the ADI-R. The CARS translation[19] was also based on the Sperber's model. The translation was performed by two independent translators, and both versions were compared and discussed by the researchers. A preliminary version was back-translated and used in the study to assess the psychometric properties.

Table 2. Information on the use and authorship of the instruments studied

Instrument	Type of use	Age group	Administration time	Administration mode	Copyrights Statement (according to Brazilian psychometric studies)	Authors
ABC	Screening	Over 2 years old	15 minutes	Inventory consisting of 57 items and answered by teachers or healthcare professionals	Not reported by the authors of the psychometric studies	**Original publication:** Krug et al.[29] **Psychometric studies:** Marteleto & Pedromônico[24] Marteleto et al.[16]
ADI-R	Diagnostic	Over 2 years old	From 1 hour and 30 minutes to 2 hours and 30 minutes	Interview composed of 93 items and administered to parents or caretakers	Study registration in the Western Psychological Services (WPS), owner of the ADI-R's copyright (Becker et al.[14])	**Original publication:** Lord et al.[30] **Psychometric studies:** Aguiar*[28] Becker et al.[14]
ASQ	Screening	Over 4 years old	Less than 10 minutes	Questionnaire composed of 40 items and answered by parents	Not reported by the authors of the psychometric study	**Original publication:** Berument et al.[31] **Psychometric study:** Sato et al.[18]
ATA	Screening	Over 2 years old	From 20 to 30 minutes	Coding of 23 subscales based on direct observation of the child	Not reported by the authors of the psychometric study	**Original publication:** Ballabriga et al.[32] **Psychometric studies:** Assumpção Jr. et al.[17] Assumpção Jr. et al.[25]
CARS	Diagnostic	Over 2 years old	From 5 to 10 minutes	Coding of a scale composed by 15 items based on direct observation of the child	Study registration in the Western Psychological Services (WPS), owner of the CARS's copyright (Pereira et al.[19])	**Original publication:** Schopler et al.[33] **Psychometric studies:** Matteo et al.[26] Pereira et al.[19]
M-CHAT	Screening	From 16 to 30 months old	From 5 to 10 minutes	Inventory consisting of 23 items and answered by healthcare professionals	Study conducted with authorization of the original authors (Losapio & Pondé[21]) and instrument available for *download* (Castro-Souza[15])	**Original publication:** Robins et al.[35] **Psychometric studies:** Castro-Souza[15] Losapio e Pondé[21]

ABC: Autism Behavior Checklist; ADI-R: Autism Diagnostic Interview-Revised; ATA: Autistic Traits Assessment Scale; ASQ: Autism Screening Questionnaire; CARS: Childhood Autism Rating Scale; M-CHAT: Modified Checklist for Autism in Toddlers. * The content of this thesis could not be fully accessed.

Regarding the adaptation of the ABC[24], the translation and back-translation steps and a pilot study with six mothers were mentioned. The study also reported that there were problems in understanding the 15 items and, for this reason, these items were adapted. As for the ATA[25], the authors indicated that after the translation, corrections were made by an expert, but they did not mention any pilot study on the understanding of the items. Finally, the ASQ[18] was translated, back-translated and evaluated by experts and a committee that considered the semantic, cultural and idiomatic equivalences.

Reliability of the instruments

In general, internal consistency was the method most commonly used in the analyzed studies[14,15,17-19], primarily using Cronbach's alpha. For the ASQ[18], the reliability of which was also investigated, the value of Cronbach's alpha for the subscales ranged from 0.63 to 0.84 and it was 0.89 for the overall score[18]. Furthermore, the KR-20 was performed, which showed very similar values to the Cronbach's alpha. The reliability of the ASQ was also investigated by the retest of part of the sample, approximately six months after the first administration. The authors calculated the Kappa for each item,

which showed that nine of the forty items had a value below 0.60. In addition, five items showed low classification power.

Regarding CARS[19], the Cronbach's alpha for the overall score was adequate (0.82) and a retest was performed on part of the sample at least four weeks after the first administration. The Kappa coefficient was 0.90. The assessment of the M-CHAT's reliability was conducted by Castro-Souza[15] using the translation proposed by Losapio and Pondé[21], and it showed a satisfactory Cronbach's alpha (0.95) for the overall score of the 20 items of the scale. Internal consistency was lower for the ATA, being 0.71 for the overall score of its 23 items[17].

In the administration of the ADI-R[14], the interviewers did not know the participants' diagnoses and, subsequently, inter-rater reliability was determined. It was not possible to calculate Kappa for one of the items, another had a moderate value and the rest indicated values ranging from substantial to almost perfect. The average Kappa was 0.82, considered being almost perfect, and Cronbach's alpha was satisfactory (0.96).

Inter-rater reliability was also investigated for the ABC[16]. The authors compared the responses of the mothers of children with ASD with the responses of the professionals that monitored these children. The agreement between the

groups was low both in relation to the overall score of the inventory and to its subareas.

Validity of the instruments

The evidence for the validity of the M-CHAT was investigated by means of exploratory factor analysis (EFA) using the principal components method and direct oblimin rotation[15]. The KMO (Kaiser-Meyer-Olkin Measure of Sampling Adequacy) was adequate (0.95), and the Kaiser method suggested four dimensions, while the parallel analysis suggested two. The author chose to extract only one factor. Three items therefore had factor loadings lower than 0.30 and the remainder showed values between 0.40 and 0.84.

The study of the ASQ[18] showed evidence of criterion validity, demonstrating that the overall score of the instrument was significantly higher for the group with ASD than for the groups with Down syndrome and with Psychiatric Disorders. Thus, the cutoff value of 14.5 indicated good sensitivity (92.5%) and specificity (95.0%)[18], similar to the original study[31].

The study of the ADI-R sought evidence of criterion validity by comparing the results of the group with ASD to those of the group with ID[14]. The data indicated that the overall score, the score per domain and 42 items were able to discriminate between the groups. In addition, the instrument showed a sensitivity and specificity of 100%. No information was provided about the cutoff values.

The validity of the ATA[17] was studied by comparing individuals with ID and individuals with autism. The former obtained a mean overall score of 15.76 points and the latter of 31.56. There was no reported analysis of the difference between these scores. The authors also noted that there was poor agreement between the DSM-IV criteria, showing a Kappa of 0.04. Moreover, the cutoff value used was 15 points with a sensitivity of 0.96[17]. A complementary study of the ATA validity was conducted in 2008[25] comparing the same diagnoses from the previous study. The autism group had a significantly higher score (30.49) in the ATA than the ID group (14.92). The scale had a sensitivity of 0.82 and specificity of 0.75, increasing the cutoff value to 23 points.

The ABC[24], with a cutoff value of 67/68 as suggested by the original authors of the instrument, obtained lower sensitivity (57.89%) and specificity (94.73%) values. When the cutoff value was reduced to 48/49, the values were more than adequate (92.6% and 92.6%, respectively). The study compared the responses of the mothers of children diagnosed with autistic disorder, language disorders and children without complaints regarding linguistic and social impairment. The children with autism had a significantly higher overall inventory score[24].

The association of CARS[19] with the ATA and the Global Assessment of Functioning (GAF) Scale was investigated, the latter being a subjective assessment contained in the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV), to compose Axis V of the multiaxial assessment[20]. A strong positive association was found with the ATA (r = 0.89) and a moderate and negative relationship was found with the GAF (-0.75). Another study sought evidence of the CARS' validity[26]. The responses of mothers of children with autism and ID were compared. The overall score on CARS was significantly higher for the group with autism, and increasing the cutoff value from 30 to 33 points led to a sensitivity of 0.81 and specificity of 0.83. Children with autism demonstrated a significantly higher score (40.38) than other children (26.38) on CARS.

In general, most of the studies analyzed sought to assess whether the instrument in question could differentiate the ASD group from other control groups[14,18,24-26]. Sensitivity and specificity analysis was also widely used[14,18,24-26], primarily using the ROC curve. Furthermore, one study evaluated the evidence of validity using EFA[15] and another investigated the association of the focal instrument of the study with others[19], although psychometrically robust instruments were not used.

DISCUSSION

Considerations on the sample used

A clear understanding of the composition of the sample, its selection, inclusion and exclusion criteria and demographic profile are essential for the correct interpretation of the results obtained in the survey[13]. In general, the descriptions of the participants in the assessed studies lacked detail, as only one study[14] described its sample exclusion criteria, albeit partially.

There was also a lack of information regarding the participants' diagnostic assessment, whether ASD, ID or another psychiatric disorder. The intelligence assessment is especially important for a correct analysis of the data, as a measure of extremely low IQ is one of the diagnostic criteria for the ID, as stated in the Diagnostic and Statistical Manual of Mental Disorders[20]. No study reported the IQ of the participants with ASD, despite the fact that such data are important due to the frequent association between ID and the disorder[3]. More specifically, the severity of ASD symptoms such as the quality of socio-communicative behavior is influenced by IQ.

In fact, Brazil lacks appropriate instruments for assessing individuals with ASD, especially preschool children. As most of these children have language disorders, instruments involving the use of pencil and paper or the understanding of instructions are, in general, difficult to administer. This feature requires the use of intelligence assessment tools that are appealing and meaningful to these children, involving the manipulation of concrete objects rather than just images. Moreover, the motivation to interact with the assessor is

often reduced in children with ASD, especially in situations of cognitive demand. Instruments such as the Merrill-Palmer-Revised (Scales of Development) and The New Reynell Developmental Language Scales may be examples of appropriate instruments for this purpose, although currently they are not available for use in Brazil.

Adaptation process

Adaptation is a broad and complex process that encompasses the translation step[9,10]. Although most studies perform a translation followed by a back-translation and then initiate procedures for research on semantic, idiomatic and cultural equivalences, Borsa *et al.*[9] recommend that the back-translation should be the last step before the pilot study, so that the original author of the instrument can assess possible cultural changes. Furthermore, a cross-cultural adaptation of the instrument can be initiated with a study of the conceptual equivalence of the items even before the first translation[10].

In this regard, some studies did not address the process of adaptation in detail, i.e., the steps of translation, who the translators were, what changes were made to each item, whether there was a pilot study, how the items were understood by the target population and whether the original author evaluated the final version. In part, this limitation is due to publication bias because currently, most scientific journals in the field do not accept articles that describe this procedure step-by-step as their primary objective. This limitation becomes a major problem for readers who do not have the opportunity to critically evaluate and judge the coherence and robustness of the adaptation. The adaptation of the M-CHAT met most of the recommendations of the model proposed by Borsa *et al.*[9], with the exception of the use of independent translators and the assessment of the validity of the final version. The other studies only briefly discussed how the adaptation was carried out.

The adaptation of the ADI-R was briefly described so that was not possible to know which changes were made and also whether the items were easily understood by the target population. In this regard, it is important to note that the ADI-R should be administered only by trained professionals at accredited centers, which reduces the margin of error in relation to the understanding of the items and coding of the results.

The ABC adaptation process was also presented briefly[24], but as the paper presents the final version of the instrument, it is possible to verify its items. Some items did appear to be ambiguous, giving rise to different interpretations. Some translated examples may be cited: "uses toys inappropriately", "lacks a social smile", "repeats sequences of complicated behaviors (covering things, for example)" and "uses more than 15 and less than 30 phrases daily to communicate" (Marteleto and Pedromônico[24], p. 298). Some essential points could therefore be clarified to provide a greater

understanding of the ABC adaptation process. For example, the study did not address whether independent translations were made, whether there was expert assessment, how the changes to be performed on items were defined, what these changes were and what the opinion of the author of the original instrument was. Thus, some fundamental steps of Borsa *et al.*'s model[9] were not addressed. This issue will be considered again in the section on the reliability of the instruments.

Another study that attached the instrument to the article and enabled some questions about the wording of the items was the adaptation of the ATA[25]. Again, the ambiguity of some translated items may elicit different interpretations, *e.g.*, "If the adult does not respond to his/her demands, the child acts by interfering in the conduct of that adult", "Adheres to a time sequence (Everything in its time)" and "When following stimuli with the eyes only does so intermittently" (Assumpção Jr. *et al.*[25], p. 25).

The ASQ study[18], like most of the highlighted studies, did not report whether there were independent translations, whether the understanding of items was investigated in a pilot study or whether the original author evaluated the adaptation. Regarding CARS[19], there was no reported pilot study to assess understanding of the items by the target population.

In relation to the other instruments, a concern with performing their back-translation may be observed, but the same attention is not given to how the translation is performed (independent versions), to contacting the original author and especially to evaluating understanding of the items. This latter point is particularly important because the lack of understanding of the items directly affects the reliability and validity of the tool[27]. However, a limitation on the full and detailed presentation of adaptation procedures appears to be a restriction for the publication of such results imposed by the scientific journals except in the cases where the study includes the investigation of psychometric properties such as reliability and validity.

Reliability of the instruments

The methods most commonly used to investigate the reliability of an instrument are test-retest, inter-rater agreement, the comparison of forms and internal consistency[27]. Concerning the internal consistency, although there is no consensus on the minimum acceptable value, values below 0.60 are considered to be inadmissible[27].

The ABC[16] did not show a satisfactory inter-rater agreement (mother of children with ASD and professionals). One aspect to be considered in this study is in relation to the understanding of the items by the mothers. As observed previously in the subsection relating to adaptation, some ABC items may be considered to be ambiguous, giving rise to different interpretations. The authors stated that the mothers,

unlike the participating professionals, responded to the inventory in an interview conducted by trained interviewers. This procedure was adopted to minimize the influence of education level. However, the lack of standardization in the application of the instrument can also influence the results regarding its reliability[27].

The ASQ study[18] reported satisfactory values of internal consistency (Cronbach's alpha and KR-20). The Cronbach's alpha for the overall score had a higher value than that for the questionnaire subscales, possibly because the internal consistency value is influenced by the number of items in the instrument[8]. Furthermore, the instrument did not show good temporal stability or classification power. On the other hand, CARS[19], ADI-R[14], M-CHAT[15] and ATA[17] showed better results in terms of reliability.

Note that there is no single reliability because each method attempts to address the types of error that can affect it, such as failures in correcting the test, in its content, in administration conditions (standardization) or the personal circumstances of the respondent[27]. One way to consider all of these sources of error is using Generalizability Theory, which is based on an analysis of variance[8,27]. However, none of the reviewed studies considered this theory.

Validity of the instruments

It is currently expected that the assessment instruments show evidence of validity that can be aggregated across different studies. No single study exhausts the possibilities for the investigations of validity[8]. The EFA is widely used and appropriate for investigating the latent structure of an instrument, thus providing evidence of construct validity[8]. Nevertheless, regarding the analyzed studies, this technique was used only in the M-CHAT[15] validity study. However, as the scale is answered dichotomously (yes/no), it would be more prudent to use an EFA analysis with tetrachoric correlations, which are used specifically for categorical variables.

Note that ASD assessment instruments comprising items for which the response is dichotomous (yes/no) are common. However, ASD is a complex disorder and its diagnosis lies in the quality, not necessarily in the absence or presence, of particular behaviors. It is therefore critical that attention is paid to the dichotomous response items because these may not be as accurate for assessing particular behavior as those assessment items investigating quality, that is, the way the child behaves.

Good evidences of validity were found regarding ASQ[18], ABC[24] and specially ADI-R[14] that showed a sensitivity and specificity of 100%. On the other hand, the first ATA validity study[17] showed poorer results once the agreement with the DSM-IV criteria was low, although the sensitivity value was satisfactory. Furthermore, the authors did not investigate if the ATA scores were significantly different among groups. The second study involving the ATA[25] showed significant differences among groups concerning the scores, but the sensitivity was lower than the value found in the first study, although still satisfactory.

One of the validity studies regarding CARS[19] was performed by comparing it with instruments that are not well-established in terms of psychometric properties. It is important to note that the area of psychometrics recommends that gold standard instruments should be used for this purpose. ATA and CARS are similar in terms of validity evidences and no GAF validity and reliability studies have been conducted on the Brazilian population. Although GAF is part of the DSM-IV, there appears to be no evidence of its psychometric quality that would justify its use in validation studies of other instruments. Another study regarding CARS[26] showed better evidences of validity. Furthermore, the original version of CARS also offers distinct scores for different levels of severity of the assessed symptoms. None of the studies assessed the adequacy of these severity levels for the Brazilian population. Pereira *et al.*[19] reported this classification in their sample, but without a comparative standard.

Ethical considerations

The International Test Commission (ITC)[34] drafted guidelines for the use of tests including important ethical issues that were briefly addressed in the present study. The guidelines point out the responsibility of the evaluator in choosing instruments that have evidence of validity for its purposes, as well as information about the scales that may be more appropriate for each situation. The analysis presented in this study aimed to provide information to assist clinicians and researchers in selecting their assessment tools.

It should be noted that the use of instruments in both the clinical and research settings entails considering the existence of copyright, the legal function of which is to protect the authors' intellectual property, among other things. Based on the instruments discussed in this study, it is common for international publishers to manage the copyright of such instruments. It is often necessary for psychometric studies to be conducted under the publisher's formal authorization by registering the project. Even instruments with existing validation studies may not be released for use in either the clinical environment or research projects other than those registered with the publisher. This release is subject to the purchase of the instrument's copyright by Brazilian publishers. Thus, the lack of explicit information on copyrights in some of the reviewed studies involves risk regarding the illegal use of these instruments.

CONCLUSION

The present study identified six ASD assessment instruments that are being studied for the Brazilian population. In gene-

ral, the adaptation process of the instruments was briefly described in the revised studies. In terms of reliability, all of the instruments that assessed internal consistency showed adequate values. In addition, the ADI-R and the CARS adaptations also satisfactorily contemplated inter-rater reliability and test-retest indices, respectively. Finally, all studies aiming to validate instruments showed evidence of validity and sensitivity. However, the lack of information on copyrights in some of the studies compromises the appropriate use of instruments.

Thus, based on this review and considering both the psychometric aspects and the copyright information, the screening instrument that currently appears to be best indicated for clinical and research use is the M-CHAT[15,21], although further evidence of validity is needed. This instrument is copyrighted but was made available for free use by the original authors[35]. Based on this review it was also noticed that there are still no specific ASD diagnostic tools available for use in Brazil. This lack of diagnostic instruments consists in a critical situation for the improvement of clinical practice and the development of research in this area. Although screening instruments are important, they do not replace those used for diagnosis in any way.

An important limitation of the present study is the possibility that the descriptors used and the databases on which the searches were performed did not cover all of the studies investigating the psychometric properties of ASD assessment instruments in Brazil. Furthermore, it was decided to only include international ASD assessment instruments, and instruments developed in Brazil were not included in this review. It is therefore not expected that this study has exhausted all papers published on this subject.

This type of research is critical for several reasons. First, the screening and diagnosis of ASD needs to be performed using quality instruments, with satisfactory psychometric properties. Furthermore, the accuracy of the study results in the field of ASD is directly related to the psychometric quality of the instruments used. This accuracy ensures safety in decision-making related to clinical interventions, community programs and public policies in the field in question. Finally, studies such as this can provide brief guidelines about care and responsibility in the use of assessment instruments, in this case, specifically in the field of ASD. The methodological quality of the studies, the careful description of the procedures adopted and the information about copyright must be considered carefully before using an instrument. Furthermore, the copyright requirements have to be described when publishing the Brazilian version of some instruments as for many cases the permission for their use is restricted to the validity study. In other words, the validity study does not imply further free access to the Brazilian version of those instruments.

INDIVIDUAL CONTRIBUTIONS

CONFLICT OF INTERESTS

REFERENCES

1. American Psychiatric Association (APA). DSM-5 Diagnostic and Statistical Manual of Mental Disorders. 5th ed. Arlington (VA): American Psychiatric Publishing; 2013.

2. Freitag CM. The genetics of autistic disorder and its clinical relevance: a review of the literature. Mol Psychiatry. 2007;12(1):2-22.

3. Rutter ML. Progress in understanding autism: 2007-2010. J Autism Dev Disord. 2011;41(4):395-404.

4. Fombonne E. Epidemiology of pervasive developmental disorders. Pediatr Res. 2009;65(6):591-8.

5. Centers for Disease Control and Prevention. Prevalence of autism spectrum disorders: autism and developmental disabilities monitoring network. MMWR. 2012;61(3):1-19.

6. Newschaffer CJ, Croen LA, Daniels J, Giarelli E, Grether JK, Levy SE, et al. The epidemiology of autism spectrum disorders. Annu Rev Public Health. 2007;28:235-58.

7. Charman T, Gotham K. Measurement issues: screening and diagnostic instruments for autism spectrum disorders – Lessons from research and practice. CAMH. 2013;18(1):52-63.

8. Urbina S. Fundamentos da testagem psicológica. 1ª ed. Porto Alegre: Artmed; 2007.

9. Borsa JC, Damásio BF, Bandeira DR. Adaptação e validação de instrumentos psicológicos entre culturas: algumas considerações. Paideia. 2012;22(53):423-32.

10. Gjersing L, Caplehorn JRM, Clausen T. Cross-cultural adaptation of research instruments: language, setting, time and statistical considerations. BMC Med Res Meth. 2010;10(13):1-10.

11. Pasquali L. Validade dos testes psicológicos: Será possível reencontrar o caminho? Psic Teor Pesq. 2007;23(spe):99-107.

12. American Educational Research Association, American Psychological Association & National Council on Measurement in Education. Standards for educational and psychological testing. 2nd ed. Washington: American Educational Research Association Publications; 1999.

13. Sabadini AAZP, Sampaio MIC, Koller SH. Publicar em psicologia: um enfoque para a revista científica. São Paulo: Associação Brasileira de Editores Científicos de Psicologia; 2009.

14. Becker MM, Wagner MB, Bosa CA, Schmidt C, Longo D, Papaleo C, et al. Translation and Validation of Autism Diagnostic Interview-Revised (ADI-R) for autism diagnosis in Brazil. Arq Neuropsiquiatr. 2012;70(3):185-90.

15. Castro-Souza RM. Adaptação brasileira do M-CHAT (Modified Checklist for Autism in Toddlers) [tese]. Brasília (DF): Universidade de Brasília; 2011.

16. Marteleto MRF, Menezes CGL, Tamanaha C, Chiari BM, Perissinoto J. Aplicação do Inventário de Comportamentos Autísticos: a concordância nas observações entre pais e profissionais em dois contextos de intervenção. Rev Bras Psiquiatr. 2008;30(3):203-8.

17. Assumpção Jr. FB, Kuczynsk E, Gabriel MG, Rocca C. Escala de avaliação de traços autísticos (ATA): validade e confiabilidade de uma escala para a detecção de condutas artísticas. Arq Neuropsiquiatr. 1999;57(1):23-9.

18. Sato FP, Paula CS, Lowenthal R, Nakano EY, Schwartzman DBJS, Mercadante MT. Instrument to screen cases of pervasive developmental disorder: a preliminary indication of validity. Rev Bras Psiquiatr. 2009;31(1):30-3.

19. Pereira A, Riesgo RS, Wagner MB. Autismo infantil: a validação da Childhood Autism Rating Scale para uso no Brasil. J Pediatr. 2008;84(6):487-94.

20. American Psychiatric Association (APA). DSM-IV-TR: Manual diagnóstico e estatístico de transtornos mentais. 4th ed. Porto Alegre: Artes Médicas; 2002.

21. Losapio MF, Pondé MP. Tradução para o português da escala M-CHAT para rastreamento precoce de autismo. Rev Psiquiatr Rio Gd Sul. 2008;30(3):221-9.

22. Reichenheim ME, Moraes CL. Operacionalização de adaptação transcultural de instrumentos de aferição usados em epidemiologia. Rev Saúde Pública. 2007;41(4):665-73.

23. Sperber, AD. Translation and validation of study instruments for cross-cultural research. Gastroenterology. 2004;126:S124-8.

24. Marteleto MRF, Pedromônico MRM. Validity of Autism Behavior Checklist (ABC): preliminary study. Rev Bras Psiquiatr. 2005;27(4):295-301.

25. Assumpção Jr. FB, Gonçalves JM, Cuccolichio S, Amorim LCD, Rego F, Gomes C, et al. Escala de Avaliação de Traços Autísticos (ATA): segundo estudo de validade. Med Reabil. 2008;27(2):41-4.

26. Matteo J, Cucolicchio S, Pacheco R, Gomes C, Simone MF, Assumpção Jr. FB. Childhood Autism Rating Scale (CARS): um estudo de validade. Med Reabil. 2009;28(2):34-7.

27. Hogan TP. Introdução à prática de testes psicológicos. 1st ed. Rio de Janeiro: LTC; 2006.

28. Aguiar CLC. A tradução da ADI-R, Entrevista Diagnóstica de Autismo-Revisada, "Autism Diagnostic Interview" [tese]. São Paulo (SP): Universidade Presbiteriana Mackenzie; 2005.

29. Krug D, Arick J, Almond P. Autism Screening Instrument for Educational Planning: ASIEP-2. 2nd ed. Texas: Pro-Ed; 1993.

30. Lord C, Rutter M, Le Couteur A. Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. J Autism Dev Disord. 1994;24(5):659-85.

31. Berument SK, Rutter M, Lord C, Pickles A, Bailey A. Autism screening questionnaire: diagnostic validity. Br J Psychiatry. 1999;175:444-51.

32. Ballabriga MCJ, Escudé RMC, Llaberia ED. Escala d'avaluació dels trets autistes (A.T.A.): validez y fiabilidad de una escala para el examen de las conductas autistas. Rev Psiquiatr Infanto-Juv. 1994;4:254-63.

33. Schopler E, Reichler R, Renner BR. The Childhood Autism Rating Scale (CARS). 10th ed. Los Angeles: Western Psychological Services; 1988.

34. International Test Commission (ITC). Guidelines on test use. Int J Test. 2001;1:93-114.

35. Robins DL, Fein D, Barton ML, Green JA. The Modified Checklist for Autism in Toddlers: an initial study investigating the early detection of autism and pervasive developmental disorders. J Autism Dev Disord. 2001;31(2):131-44.