

Marizete Ilha Ceron¹
Marileida Barichello Gubiani¹
Camila Rosa de Oliveira²
Márcia Keske-Soares¹

Keywords

Speech
Speech Articulation Tests
Articulation Disorders
Assessment
Child

Descritores

Fala
Testes de Articulação da Fala
Transtornos da Articulação
Avaliação
Criança

Corresponding author:

Marizete Ilha Ceron
Av. Roraima, 1000, Prédio 26,
CCS, 4º andar, Sala 1418, Bairro
Camobi, Santa Maria (RS), Brasil,
CEP: 97105-900.
E-mail: marizeteceron@hotmail.com

Received: October 23, 2016

Accepted: October 23, 2017

Evidence of validity and reliability of a phonological assessment tool

Evidências de validade e fidedignidade de um instrumento de avaliação fonológica

ABSTRACT

Purpose: To present evidence of the validity and reliability of a phonological assessment tool developed to assess the phonological inventory of Brazilian Portuguese. **Methods:** The study included 866 children aged between 3 and 8:11 years, divided into three groups: typical, control and clinical. Participants were evaluated using a phonological assessment software, which prompted the spontaneous naming of a series of images. The children's responses were audio recorded and transcribed at the time of the assessment, by the software itself. The Cronbach's alpha coefficient was used to evaluate the internal consistency of the instrument for reliability and validity purposes. Criterion validity was examined by comparing the performance of different groups using Student's t-test for independent samples. Intra- and inter-rater agreement were investigated using Kendall's tau. Results were considered significant at $p \leq 0.05$. **Results:** The present study provided evidence of validity and reliability (internal consistency) for this phonological assessment tool, confirming the reliability of its items and demonstrating excellent agreement rates between examiners regarding its scoring (intra- and inter-rater reliability). The criterion validity assessment demonstrated that the control group outperformed the clinical group across all phonemes, showing that test scores were successful in identifying children with speech sound disorders (phonological disorders). **Conclusion:** The present findings provide strong evidence of the validity and reliability of this phonological assessment tool.

RESUMO

Objetivo: Apresentar evidências de validade e fidedignidade de um instrumento de avaliação fonológica (INFONO) desenvolvido para avaliar os fonemas do Português Brasileiro. **Método:** Participaram do estudo 866 crianças com idades entre 3 e 8:11 anos, divididas em grupos: típico, controle e clínico. Os participantes foram avaliados pelo INFONO por nomeação espontânea. A produção da criança foi gravada e transcrita no momento da avaliação, no próprio *software*. Para análise de validade e fidedignidade, foram analisadas a consistência interna a partir da técnica *Alpha de Cronbach*. Para a validade de critério, comparou-se o desempenho entre os grupos através do teste *t* de Student para amostras independentes. A fidedignidade foi analisada pela concordância intra e interavaliadores por meio do Teste de Kendall. Considerou-se significância quando $p \leq 0,05$. **Resultados:** O INFONO apresentou evidências de validade e fidedignidade (consistência interna), indicando uma confiabilidade satisfatória dos itens, bem como excelente concordância entre os avaliadores em relação aos escores do teste (confiabilidade intra e interavaliador). Em relação à validade de critério, o desempenho do grupo clínico para todos os fonemas foi inferior ao grupo controle, mostrando que os escores são sensíveis para identificar crianças com desvio fonológico. **Conclusão:** O INFONO apresentou fortes evidências de validade e fidedignidade.

Study conducted at Programa de Pós-graduação em Distúrbios da Comunicação Humana, Universidade Federal de Santa Maria – UFSM - Santa Maria (RS), Brasil.

¹Departamento de Fonoaudiologia, Universidade Federal de Santa Maria – UFSM - Santa Maria (RS), Brasil.

²Faculdade Meridional – IMED - Passo Fundo (RS), Brasil.

Financial support: Programa Pesquisador Gaúcho (PqG) - Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS), project number 11/2184-8.

Conflict of interests: nothing to declare.



This is an Open Access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

INTRODUCTION

Speech sound disorders can be caused by various etiologies, and lead to impairments at several levels of speech production, including linguistics/phonology and/or motor planning⁽¹⁻⁴⁾. Speech pathologists face the constant challenge of correctly diagnosing and distinguishing between these conditions in order to plan an effective intervention.

The present study will discuss the phonological aspects of speech sound disorders, known as phonological deviations, which are among the most prevalent alterations in children^(4,5) and a major focus of scientific research^(4,6,7). Phonological deviations are characterized by the linguistic disorganization of the phoneme inventory. Children with these conditions may omit or replace certain phonemes, especially consonants and consonant clusters, at an age where such behaviors should no longer occur^(8,9).

Phonological deviations can be evaluated and diagnosed using assessment instruments such as the *Goldman Fristoe 2 – Test of Articulation – GFTA 2*⁽¹⁰⁾ and the *Clinical Assessment of Articulation and Phonology – CAAP*⁽¹¹⁾. In Brazil, the instruments most frequently used for this purpose are the Children's Phonological Assessment (Avaliação Fonológica da Criança – AFC)⁽¹²⁾ and the Child Language Test - Phonology (ABFW - Teste de Linguagem Infantil – Fonologia)⁽¹³⁾. Both instruments are elaborate and quite comprehensive, and available for use in both clinical practice and research settings in speech pathology. However, neither has undergone psychometric evaluation in order to examine their validity and reliability in the assessment of Brazilian Portuguese (BP)-speaking children. The importance of assessing the validity and reliability of diagnostic instruments is emphasized in the international literature^(14,15). In fact, such instruments should only be made available for general use after undergoing psychometric evaluation.

The development of any assessment instrument should ensure that it measures what it is intended to measure, and that its results reflect a given ability with no influence from any variables other than those stated in the aims of the instrument^(2,10,14,15). This can be ensured by submitting the instrument to validity and reliability testing.

Validity refers to the precision of any inferences drawn from the scores of a given instrument⁽²⁾. As such, the evidence of validity for an assessment instrument determines the extent to which it measures what it is intended to measure^(10,16). This is an especially important feature of diagnostic instruments, since the precision of scores obtained from these measures will contribute directly to diagnostic decision-making.

Like validity, reliability is crucial for test development, as it speaks to the degree to which an instrument is susceptible to different sources of error (e.g. variation between raters or poor intra-rater consistency). These issues can also be a threat to test validity^(2,17). As such, it is important to analyze the evidence of intra- and inter-rater reliability for the items in an instrument. This procedure will show the extent to which the test scores are susceptible to error.

The international literature has shown growing interest in the search for instruments with established psychometric properties

in the assessment of speech sound disorders^(2,14,15,17-19). Some of these studies^(12,13,15) discuss multiple phonological assessment instruments whose validity and reliability has already been examined. Yet this is still a recent development in Brazil, where none of the available phonological assessment instruments have undergone psychometric evaluation^(14,20,21).

This constitutes a limitation in phonological assessment in BP, which compromises the diagnosis and treatment planning for children with speech disorders, especially in the case of inexperienced examiners. Therefore, the development of assessment protocols with robust psychometric properties is crucial for clinical practice and research in speech pathology^(14,15,20). In addition to psychometric studies, information technology may also contribute to phonological assessment in speech pathology. The use of computer software for phonological assessment may be both more appealing to children, and advantageous for practitioners, as it makes for faster and simpler test administration.

In light of these observations and the growing concerns regarding the availability of valid and reliable instruments for clinical and research purposes, this study aimed to evaluate the validity and reliability of the Phonological Assessment Instrument INFONO, developed specifically to evaluate phonemes in BP.

METHODS

Participants

The sample consisted of children aged between 3 years and 8 years 11 months, attending 12 different schools (8 public and 4 private) in southern Brazil. Of the 1448 children invited, 1076 (73%) were authorized to take part in the study. Participants who were bilingual, had subjective or suspected hearing loss, neurological and/or psychological conditions, intellectual disability, a diagnosis of autism or Down's syndrome, or previous speech therapy were excluded from the sample. This information was obtained from screening questionnaires administered to parents/guardians and teachers. Children with signs of any alterations which could lead to speech impairments (e.g. anterior open bite, lisp, tongue thrust, probable language and/or vocabulary deficits) were also excluded from the study. These alterations were detected during brief informal conversations with each child, where they were asked about their school, their age, what day it was, what they liked to play, what they liked to eat, what they were currently doing in class, if they liked animals, which animals they liked, etc., or during the administration of the INFONO. These criteria led to the exclusion of 210 children from the study.

The final sample therefore comprised $n = 866$ participants, divided into two groups: a typically developing and a clinical group. The former group was composed of children with typical phonological development ($n = 733$), who were able to produce all phonemes expected for their age (e.g. a 3-year-old who could not pronounce the sound /r/ would be included in this participant group). The atypical group ($n = 133$), on the other hand, included children with atypical phonological development, who displayed omissions or substitutions beyond the age at which these phenomena were expected to cease.

Chart 1. Patient characteristics

Groups	N	Age in years and months		Gender		Type of school		
		M	SP	Female <i>n</i> (%)	Male <i>n</i> (%)	Private <i>n</i> (%)	Public <i>n</i> (%)	
Typical	733	6.08	1.58	332 (45.3)	401 (54.7)	422 (57.6)	311 (42.4)	
Atypical	133	5.87	1.23	54 (40.6)	79 (59.3)	58 (43.6)	75 (56.4)	
Control	Typical	228	5.89	1.31	100 (43.9)	128 (56.1)	104 (45.6)	124 (54.4)
	Clinical	114	5.88	1.32	50 (43.9)	64 (56.1)	52 (45.6)	62 (54.4)

Caption: N = total number of participants; *n* = number of participants; M = Mean; SD = Standard deviation

Examples include a 3-year old who showed alterations in the production of the phoneme /b/ or a 5-year-old who was unable to pronounce the /r/ sound. This classification was carried out by the examining speech pathologist, based on her clinical experience in assessment and treatment, and on the national literature^(22,23) on phonological acquisition.

A third group was then formed by both typically developing and atypical participants, which was referred to as the control group. These individuals were drawn from the typically developing (*n* = 228) and clinical groups (*n* = 114), and matched by age group, type of school and gender. This group comprised a total of 342 children (clinical and control groups). The clinical and control groups did not significantly differ in terms of their mean age ($F = 0.000$; $p = 0.977$), gender distribution ($X^2 = 0.000$; $p = 1.000$) or school type ($X^2 = 0.000$; $p = 1.000$). Sample characteristics are described in Chart 1.

Instrument

The INFONO is a software package developed for assessment purposes, to be used by speech pathologists with the help of a computer. The instrument was developed in four stages: literature review/stimulus selection⁽²⁴⁾; analysis by expert panels⁽²⁴⁾; analysis by non-expert users⁽²⁴⁾; and pilot study⁽²⁵⁾. These stages were crucial for the development of the INFONO software. The developmental process has already been completed, and the INFONO is now undergoing final psychometric testing before being made available for the general use of speech therapists. The international literature^(14,15) has suggested that psychometric studies are crucial for all assessment instruments involved in diagnostic decision-making and should only be made publicly available after these studies are completed.

The INFONO software allows for data collection in the form of audio recordings for later review and transcription. Data was collected via spontaneous naming of test stimuli. After each item, the examiner must select the transcription which corresponds to the child's response from a list of options provided by the software. After the evaluation is complete, these selections can be reviewed and compared to the audio recordings. The list of possible transcriptions was developed based on the phonological processes observed in children with both typical development and phonological deviations. If none of the options corresponds to the child's response, the examiner

can enter a new transcription into a keyboard provided by the software, noting any distortions present in the child's speech.

The test stimuli for spontaneous naming consist of 84 colorful animated drawings in the form of gifs (animations formed by merging multiple *GIF* images into a single file, creating the illusion of movement), representing the target words. The examiner is also given a prompt question which they can ask the child during the assessment to facilitate the identification and production of the target word, such as: "*He uses the pencil to...?*" (write), "*What animal is this?*" (dog), etc. Other questions can also be used to facilitate the production of the target-word. It is important that the child produces all target words in the test, since they contain all phonemes of Brazilian Portuguese in all possible word and syllable positions.

After completing the assessment and transcriptions, the examiner clicks on a button to generate the results. The *software* then provides the following data: analysis of test results (list of target words and phonetic transcriptions of the child's responses), contrast analysis (number of correct answers, omissions and substitutions for every phoneme in Brazilian Portuguese) in order to describe the child's phonetic (presence or absence of different phonemes) and phonological inventory (percentage of errors, omissions and substitutions for each phoneme and onset clusters [OC]); analysis of distinguishing features; analysis of phonological processes; and severity of the phonological disorder.

The instrument does not evaluate vowels, since these are acquired early in development and are much less likely to be affected in Brazilian Portuguese speakers with phonological disorders.

Procedures

This study was conducted according to all relevant ethical guidelines, including approval by the Research Ethics Committee of the Universidade Federal de Santa Maria (UFSM), under protocol number 23081.005433/2011-65. All parents and guardians provided written consent to their children's participation in the study. The children, in turn, were only evaluated after assenting to participate in the assessment.

Firstly, a brief conversation was carried out with each child to detect any impairments which may justify their exclusion from the study. Both this procedure and the administration of the INFONO were conducted by three doctoral students in speech pathology as well as a speech pathologist with over

10 years' experience in this area of study. All examiners had previous training on how to administer and use the INFONO in a research context.

After the initial conversation, participants were administered the INFONO. Data was collected in individual testing sessions conducted in a location provided by the children's schools. The administration of the INFONO takes approximately 20 minutes and was carried out according to recommended standard procedures. Children's responses to the test were recorded and transcribed during the assessment using the software itself. The data provided by the software regarding the phonological inventory of each child was later used to create a database in SPSS where psychometric analyses were performed in order to investigate the validity and reliability of the scores provided by the INFONO, as follows:

(a) Validity

The precision of INFONO scores was determined based on its internal consistency and criterion validity (i.e. distinction between clinical and control groups). Internal consistency refers to the extent to which the test items reflect the intended purpose of the instrument⁽¹⁴⁾. Criterion validity speaks to the effectiveness of the instrument at describing the performance of a specific group of individuals⁽¹⁹⁾. In the present study, internal consistency was calculated using scores from the typically developing group ($n = 733$), while criterion validity was examined based on the control group.

(b) Reliability

Reliability refers to the consistency of test scores, ensuring they do not change when administered by the same examiner at different points in time (intra-rater reliability) or by different raters altogether (inter-rater reliability). These procedures reflect the extent to which the results of an instrument are reliable. Three sources of reliability data were examined in the present study: internal consistency, intra-rater reliability and inter-rater reliability.

Internal consistency was calculated as described in the validity section, since consistency (or stability of measurement) is also crucial for test validity. Intra-rater reliability was evaluated by asking examiners to listen to the audio recordings of their own test sessions and manually transcribe participants' responses (without the use of the software). The results of this procedure were then analyzed independently of the first assessment. This was done for a subgroup of 77 children (approximately 10% of the sample) assessed by the same examiner. These participants were randomly selected at least a month after their original evaluation.

Similarly, between-rater reliability was analyzed using the audio recordings of 120 children (approximately 15% of the sample). These were examined by two final year undergraduate students, both of whom had experience in the area, as they worked as research assistants in the laboratory where this study was conducted. The new transcriptions and analyses were performed independently of the first evaluation.

Data analysis

Data were analyzed using SPSS, version 22, for Windows. Internal consistency, which is relevant to both validity and reliability, was measured using Cronbach's alpha. This method was chosen because participants completed a single instrument on only one occasion. *The Cronbach's alpha* coefficient describes the correlation between each test item and the remaining items on the instrument, or between the item and the total score on the test. This technique provides a measure of covariance ranging from 0 to 1, where higher scores are indicative of greater reliability. Values greater than 0.7 reflect adequate validity and reliability.

Criterion validity was evaluated by comparing the performance of typically and atypically developing children (control and clinical groups), and by discriminant analysis. Firstly, performance was compared between the clinical and control groups using Student's t-test for independent samples. Then, a stepwise discriminant analysis using Wilks' lambda was conducted to verify the ability of the INFONO to differentiate between typically and atypically developing children. Assumptions of normality and homogeneity of variance-covariance matrices were examined using the Shapiro-Wilk and Box's M tests, respectively.

The reliability of the INFONO was also evaluated using intra- and inter-rater agreement as measured by Kendall's Tau, where values greater than 0.6 suggest adequate agreement, and therefore, satisfactory reliability. Intra- and interrater reliability was calculated by dividing phonemes according to the following positions: General onset (GO); initial onset (IO); medial onset (MO); general coda (GC); medial coda (MC); final coda (FC). Results were considered significant at $p \leq 0.05$.

RESULTS

The scores provided by the INFONO were examined for evidence of validity and reliability. The results of the internal consistency analysis are shown in Table 1, which contains the Cronbach's alpha coefficients for each phoneme and age group assessed by the instrument.

These findings showed high internal consistency (>0.7) across all age groups, with values ranging from 0.713 to 0.922 (median = 0.816). These results indicate that the items used to evaluate each phoneme in BP have acceptable validity and reliability within each age group.

Findings pertaining to intra-rater reliability are shown in Table 2, which displays the agreement rates for correct production of phonemes and consonant clusters in the INFONO.

Intra-rater reliability for these scores ranged from 0.622 to 1.0. Most phonemes were associated with excellent agreement rates across different word and syllable positions, with values often greater than 0.8. Agreement was considered "adequate" only for /s/ in GC, MC and FC; /k/ in GO and MO; /m/ in GO and IO; /r/ in MC; plosive+/r/ in GO and IO; and fricative+/r/ in GO, IO and MO. When separated according to syllable and word positions, mean percent agreement within raters was considered excellent, with values ranging from 0.822 to 0.944.

Results of interrater reliability analyses are shown in Table 3, which displays agreement rates for the correct production of phonemes and consonant clusters in Brazilian Portuguese as measured by the INFONO.

Inter-rater agreement ranged from 0.663 to 1.0. The lowest agreement rates were observed for /d/ in GO, /k/ in GO and MO, /m/ in GO and IO, and /l/ in GO, which were nevertheless classified as “good.” All other values were greater than 0.826, and therefore indicative of excellent interrater agreement. When examined separately for each word and syllable position,

mean percentage agreement between raters was excellent, with values of at least 0.93.

Evidence of validity for the INFONO can be found in Table 4, which compares the production of all phonemes in GO, GC and CO (means and standard deviations) between the clinical and control groups.

The results show that mean production accuracy was lower for the clinical group than for control participants across all phonemes examined. The production of individual phonemes differed significantly between the clinical and control groups in several cases, except for /p/, /d/ and /ŋ/ in GO and /n/ in GC; /p/ in IO; /b/, /d/ and /m/ in MO; /n/ in MC; and, /l/ and /s/ in FC. These findings show that INFONO scores can differentiate between children with impaired vs. typical phonological development.

In order to analyze which of these phonemes in simple onset, coda and/or OC were most useful in differentiating between the clinical and control groups, a stepwise discriminant analysis was performed. The procedure produced a discriminant function based on 18 phonemes and OC, which accounted for 100% of the variation between groups ($\Lambda = 0.494$; $\chi^2(18) = 602.403$; $p \leq 0.001$). Table 5 shows the standardized coefficients and phonemes included in the discriminant function. These phonemes made the most important contributions to the differentiation of typically and atypically developing children.

The discriminant function classification rates are shown in Chart 2. In 91.7% of cases, the discriminant function classification agreed with the original assessment, which speaks to the efficacy of the INFONO at differentiating between children with typical and atypical phonological development.

Table 1. Internal consistency of the INFONO for children of different ages

Age group	<i>n</i>	Cronbach's alpha
3:0-3:3	25	0.922
3:4-3:7	38	0.916
3:8-3:11	34	0.922
4:0-4:3	31	0.917
4:4-4:7	34	0.903
4:8-4:11	35	0.862
5:0-5:3	44	0.820
5:4-5:7	40	0.815
5:8-5:11	51	0.809
6:0-6:3	53	0.746
6:4-6:7	59	0.861
6:8-6:11	59	0.743
7:0-7:6	72	0.713
7:7-7:11	58	0.714
8:0-8:6	61	0.732
8:7-8:11	39	0.718
Median		0.816

Caption: *n* = number of participants

Table 2. Intra-rater reliability as measured by agreement on phoneme production accuracy

	Position					
	General Onset	General Coda	Initial Onset	Medial Onset	Medial Coda	Final Coda
/p/	0.811		1.000	1.000		
/b/	1.000		1.000	1.000		
/t/	0.973		1.000	0.973		
/d/	1.000		1.000	1.000		
/k/	0.702		1.000	0.702		
/g/	1.000		1.000	1.000		
/f/	1.000		1.000	1.000		
/v/	0.964		1.000	0.964		
/s/	1.000	0.690	0.935	1.000	0.631	0.678
/z/	1.000		1.000	1.000		
/ʃ/	1.000		1.000	1.000		
/ʒ/	1.000		1.000	1.000		
/ɲ/	1.000		1.000	1.000		
/m/	0.653		0.653	1.000		
/n/	1.000	0.856	1.000	1.000	0.962	1.000
/ɲ/	1.000			0.974		
/l/	0.816	1.000	1.000	0.816	1.000	1.000
/ʎ/	0.945			0.945		
/r/	0.973	0.830		0.971	0.700	0.878
Plos.+/l/	0.872		0.920	0.891		
Plos.+/r/	0.765		0.622	0.862		
Fric.+/l/	0.801		0.801			
Fric.+/r/	0.676		0.795	0.661		
Mean	0.911	0.844	0.936	0.944	0.823	0.889

Table 3. Interrater reliability as measured by agreement on phoneme production accuracy

Phonemes	Position					
	General Onset	General Coda	Initial Onset	Medial Onset	Medial Coda	Final Coda
/p/	1.000		1.000	1.000		
/b/	1.000		1.000	1.000		
/t/	0.941		1.000	1.000		
/d/	0.663		1.000	1.000		
/k/	0.772		1.000	0.704		
/g/	1.000		1.000	1.000		
/f/	1.000		1.000	1.000		
/v/	0.920		1.000	0.899		
/s/	1.000	0.909	1.000	1.000	0.946	0.863
/z/	1.000		1.000	1.000		
/ʃ/	0.996		0.997	1.000		
/ʒ/	1.000		1.000	1.000		
/ɲ/	1.000		1.000	1.000		
/m/	0.791		0.658	1.000		
/n/	1.000	0.952	1.000	1.000	1.000	1.000
/ɲ/	0.993			0.993		
/l/	0.748	1.000	1.000	0.866	1.000	1.000
/ʎ/	0.958			0.958		
/r/	0.960	0.873		0.960	0.996	0.870
Plos.+/l/	0.866		0.970	0.826		
Plos.+/r/	0.892		0.897	0.892		
Fric.+/l/	0.972		0.972			
Fric.+/r/	1.000		1.000	1.000		
Mean	0.934	0.934	0.975	0.959	0.986	0.933

Table 4. Comparison of phoneme production between clinical and control groups

		Control <i>n</i> = 228		Clinical <i>n</i> = 114		<i>F</i>	<i>p</i> -value
		<i>M</i>	<i>SP</i>	<i>M</i>	<i>DP</i>		
General Onset	/p/	8.99	0.09	8.95	0.26	21.215	0.084
	/b/	9.90	0.34	9.54	1.33	50.876	0.004
	/t/	18.73	0.67	18.12	1.80	34.793	≤ 0.001
	/d/	7.02	0.42	6.77	1.35	44.188	0.060
	/k/	17.94	0.27	16.72	3.71	81.367	≤ 0.001
	/g/	6.97	0.20	6.18	1.86	193.940	≤ 0.001
	/f/	6.98	0.17	6.64	1.23	63.277	0.004
	/v/	7.68	0.51	6.97	1.52	34.942	≤ 0.001
	/s/	8.96	0.25	8.28	1.78	111.313	≤ 0.001
	/z/	5.99	0.28	5.17	1.54	225.076	≤ 0.001
	/ʃ/	6.82	0.85	4.69	2.79	395.252	≤ 0.001
	/ʒ/	5.93	0.77	4.20	2.40	406.058	≤ 0.001
	/ɣ/	5.93	0.45	5.28	1.60	128.677	≤ 0.001
	/m/	5.96	0.21	5.86	0.40	36.231	0.016
	/n/	7.91	0.31	7.67	0.67	73.707	≤ 0.001
	/ɲ/	3.75	1.20	3.75	1.21	0.044	1.000
/l/	10.92	0.32	9.73	2.63	146.593	≤ 0.001	
/ʎ/	3.60	0.60	2.92	1.12	27.320	≤ 0.001	
/r/	7.56	1.68	3.72	3.47	241.044	≤ 0.001	
General Coda	/n/	7.89	0.74	7.70	1.05	27.162	0.087
	/l/	5.96	0.23	5.80	0.67	47.966	0.011
	/r/	6.37	1.69	3.72	2.67	74.978	≤ 0.001
Initial Onset	/s/	11.41	1.93	10.64	2.97	15.361	0.013
	/p/	3.99	0.09	3.95	0.26	21.215	0.084
	/b/	6.92	0.29	6.61	1.04	56.702	0.003
	/t/	2.99	0.11	2.93	0.29	28.152	0.045
	/d/	2.98	0.13	2.81	0.64	71.780	0.004
	/k/	11.00	0.07	10.32	2.19	87.960	≤ 0.001
	/g/	2.98	0.17	2.59	0.96	182.586	≤ 0.001
	/f/	3.00	0.00	2.89	0.53	46.051	0.023
	/v/	2.99	0.11	2.75	0.62	136.383	≤ 0.001
	/s/	3.36	0.49	3.16	0.88	9.707	0.021
	/z/	2.00	0.07	1.70	0.61	297.829	≤ 0.001
	/ʃ/	3.90	0.46	2.61	1.61	413.576	≤ 0.001
	/ʒ/	2.95	0.31	2.08	1.21	400.694	≤ 0.001
	/ɣ/	2.96	0.28	2.60	0.85	154.203	≤ 0.001
/m/	2.96	0.20	2.87	0.36	38.125	0.013	
/n/	2.93	0.27	2.77	0.50	54.132	0.002	
/l/	4.00	0.00	3.65	0.97	145.847	≤ 0.001	
Medial Onset	/p/	5.00	0.00	5.00	0.00	*	*
	/b/	2.99	0.11	2.92	0.38	23.865	0.073
	/t/	15.75	0.65	15.19	1.60	38.803	≤ 0.001
	/d/	4.01	0.37	3.92	0.82	27.029	0.256
	/k/	6.94	0.26	6.39	1.62	92.131	≤ 0.001
	/g/	3.99	0.09	3.60	1.02	157.564	≤ 0.001
	/f/	3.98	0.17	3.75	0.77	71.917	0.003
	/v/	4.69	0.48	4.22	1.05	37.185	≤ 0.001
	/s/	5.60	0.53	5.12	1.19	24.948	≤ 0.001
	/z/	4.00	0.24	3.46	1.11	199.564	≤ 0.001
	/ʃ/	2.92	0.43	2.08	1.26	315.886	≤ 0.001
	/ʒ/	2.97	0.54	2.12	1.30	263.857	≤ 0.001

Caption: M = Mean; SD = Standard deviation; *n* = number of participants; *F* = Ratio; IO = Initial onset; MO = Medial onset; * = T-test could not be conducted since standard deviations were equal to zero

Table 4. Continued...

	Control <i>n</i> = 228		Clinical <i>n</i> = 114		<i>F</i>	<i>p</i> -value	
	<i>M</i>	<i>SP</i>	<i>M</i>	<i>DP</i>			
	/χ/	2.97	0.21	2.68	0.83	110.945	≤ 0.001
	/m/	3.00	0.07	2.99	0.09	1.001	0.617
	/n/	4.98	0.13	4.89	0.31	59.123	0.004
	/ɲ/	3.75	1.20	3.75	1.21	0.044	1.000
	/l/	6.92	0.32	6.08	1.77	154.948	≤ 0.001
	/k/	3.60	0.60	2.92	1.12	27.320	≤ 0.001
	/r/	7.56	1.68	3.72	3.47	241.044	≤ 0.001
Medial Coda	/n/	5.12	0.47	5.11	0.73	13.032	0.862
	/l/	2.96	0.23	2.82	0.63	41.904	0.016
	/r/	2.74	0.75	1.54	1.35	175.508	≤ 0.001
	/s/	5.79	0.92	4.90	1.80	67.513	≤ 0.001
Final Coda	/n/	2.83	0.38	2.71	0.47	23.184	0.021
	/l/	3.00	0.00	2.98	0.13	16.785	0.158
	/r/	3.63	1.19	2.18	1.55	12.968	≤ 0.001
	/s/	5.62	1.56	5.74	1.80	1.383	0.531
Onset Clusters	Plosive +/l/	3.82	1.69	1.83	1.83	5.643	≤ 0.001
	Plosive +/r/	16.87	6.68	6.52	7.45	11.713	≤ 0.001
	Fricative + /l/	1.40	0.79	0.57	0.74	1.399	≤ 0.001
	Fricative + /r/	4.10	1.67	1.70	1.99	20.797	≤ 0.001
	Plosive +/l/ - IO	1.59	0.74	0.78	0.86	14.035	≤ 0.001
	Plosive +/l/ - MO	2.23	1.06	1.05	1.11	0.448	≤ 0.001
	Plosive +/r/ - IO	8.16	3.15	3.54	3.80	25.418	≤ 0.001
	Plosive +/r/ - MO	8.71	3.62	2.98	3.79	4.261	≤ 0.001
	Fricative + /l/ - IO	1.40	0.79	0.57	0.74	1.399	≤ 0.001
	Fricative + /r/ - IO	1.63	0.71	0.72	0.92	41.572	≤ 0.001
	Fricative + /r/ - MO	2.46	1.02	0.98	1.20	10.794	≤ 0.001

Caption: *M* = Mean; *SD* = Standard deviation; *n* = number of participants; *F* = Ratio; IO = Initial onset; MO = Medial onset; * = T-test could not be conducted since standard deviations were equal to zero

Table 5. Discriminant function coefficients

Phonemes	Discriminant function coefficients
/p/ in General Onset	-1.066
/s/ in General Onset	0.232
/ʒ/ in General Onset	0.417
/χ/ in General Onset	-0.368
/r/ in General Onset	0.179
/s/ in General Coda	-0.119
/k/ in IO	0.324
/v/ in IO	0.768
/j/ in IO	0.802
/l/ in IO	0.253
/p/ in MO	-1.784
/ʒ/ in MO	-0.330
/n/ in MO	0.610
/ɲ/ in MO	0.153
/r/ in FC	0.095
plosive + /r/	-0.172
plosive + /r/ in MO	0.484
fricative + /r/ in IO	-0.494
Constant	3.081
Eigenvalue	1.023

Caption: IO = Initial onset; MO = Medial onset; FC = Final coda

Chart 2. Results of the discriminant function analysis for the INFONO

Original classification	Predicted classification		Total
	Typical <i>n</i> (%)	Atypical <i>n</i> (%)	
Typical <i>n</i> (%)	704 (96%)	29 (4%)	733
Atypical <i>n</i> (%)	43 (32%)	90 (68%)	133

Caption: *n* = number of participants

DISCUSSION

The present study achieved its original aims and was able to demonstrate the validity and reliability of the INFONO, which has been confirmed as a useful tool for the phonological assessment of children with suspected of phonological disorders.

Validity and reliability analyses showed that the instrument has adequate internal consistency as demonstrated by its Cronbach's alpha coefficient (median = 0.816). This finding suggests that the items in the INFONO which evaluate phonemes in Brazilian Portuguese have satisfactory reliability. The internal consistency of other international instruments^(10,26) used in phonological assessment has also been evaluated using Cronbach's alpha coefficient. The internal consistency of the GFTA-2(10), for instance, ranged from 0.85 to 0.98 across genders and age groups, and is therefore considered adequate. The *Test para Avaliar Processos de Simplificação Fonológica* - TEPROSIF-R⁽²⁶⁾ and the CAAP⁽¹¹⁾, both of which are used for phonological assessment, have also shown high internal consistency (0.90) according to the literature.

A test is considered reliable when all its items or tasks provide similar measures of performance or evaluate the same domain⁽¹⁰⁾. In the present study, the internal consistency of the INFONO was not examined by age group, gender or school type, due to insufficient variability in participant scores.

The AFC⁽¹²⁾ and ABFW-Fonologia⁽¹³⁾ evaluate phonology in Brazilian Portuguese and are the most widely used instruments in Brazilian research^(7,27-29). However, the validity and reliability of these instruments has not been evaluated. Therefore, the contributions of this study to the development of an assessment software package which has been subjected to rigorous validity and reliability testing is of particular relevance to the advancement of speech pathology and the evaluation of children with suspected phonological impairments. The use of standardized instruments with evidence of both validity and reliability increases the accuracy with which the present or absence of a disorder can be determined.

There are several factors which can compromise the precision of an instrument, and as a result, influence diagnoses⁽¹⁸⁾ and treatment. As such, it is important to evaluate the validity and reliability of assessment instruments, to ensure they measure what they intend to measure, and that their scores reflect performance in the area of interest^(2,10,14,15). For instance, the results of a phonological assessment instrument should allow for the analysis of all phonemes in the target language, across all word and syllable positions, in order to determine the presence or absence of developmental impairments. The validity of an

instrument can be affected by the quality and reliability of its scores and by the skills of the examiner.

Reliability can be defined as the consistency of an instrument over time and across changes in test administration or scoring⁽¹⁹⁾. The scores obtained from a reliable instrument can be generalized from the assessment conditions to a wider range of situations. In the present study, reliability was examined using measures of intra- and interrater agreement. Intra-rater reliability was measured by comparing two distinct transcriptions of an assessment session carried out by the same examiner at two different time points. In the present study, excellent intra-rater agreement was observed for phonemes in different positions, with values ranging from 0.822 to 0.944. These findings suggest that the ratings provided by an examiner at different time points show an agreement of at least 0.822, which speaks to the reproducibility and diagnostic reliability of INFONO scores.

No studies to date have evaluated the intra-rater reliability of phonological assessment instruments. However, this is an extremely important factor for speech assessment instruments⁽¹⁵⁾, since it ensures the consistency of a measure when administered by the same practitioner across different contexts. In these situations, scores are not expected to change, since instruments should be able to produce consistent measurements regardless of when they are administered. Intra-rater agreement has been examined for an instrument used to evaluate speech apraxia, which yielded values ranging from 0.81 to 0.95⁽²⁾, not unlike those displayed by the INFONO in the present study.

The present study also revealed excellent interrater reliability for phoneme transcriptions, with agreement rates of over 0.93 for all word and syllable positions. In other words, when two examiners were asked to transcribe the same set of responses, they provided similar ratings in 93% of cases. Given the subjective nature of this assessment method, and the role of examiners' skills and experience in ensuring its precision, it was especially important to examine the interrater reliability of the INFONO in order to confirm its reproducibility and diagnostic reliability. The GFTA-2⁽¹⁰⁾ has also yielded interrater agreement rates ranging from 70 to 100%, with a mean value of 93% or greater, depending on phoneme position. The interrater agreement for the CAAP⁽¹¹⁾ has been calculated at 99%, suggesting that raters are highly consistent in their scoring of this instrument.

The validity procedures, which included a discriminant analysis between the clinical and control groups, provided strong evidence of criterion validity for the INFONO. As expected, mean scores across all phonemes were lower for the clinical than the control group, which suggests that the INFONO is sensitive to between-group differences in phonological development. Significant differences between the groups were observed for most phonemes and onset clusters, except for /p/, /d/ and /j/ in GO and /n/ in GC; /p/ in IO; /b/ and /m/ in MO; /n/ in MC; and, /l/ and /s/ in FC. It is possible that these phonemes are acquired and stabilized before the age of three, which is younger than the participants included in the present study. Previous studies^(22,23) appear to confirm this hypothesis. The order of phoneme acquisition in Brazilian Portuguese tends to be the following: plosives and nasals, followed by fricatives and finally, liquids^(23,30) and complex onsets.

Our results also revealed that the INFONO possesses a high degree of discriminability (91.7%) between typically and atypically developing children. The phonemes which contributed the most to this classification were /p/, /s/, /z/, /ʒ/ and /r/ in GO; /s/ in GC; /k/, /v/, /f/ and /l/ in IO; /p/, /ʒ/, /n/ and /ɲ/ in MO; /r/ in FC; plosive + /r/; plosive + /r/ in MO; fricative + /r/ in IO. These findings confirm that the INFONO has achieved one of its main goals: that of contributing to the differential diagnosis of phonological disorders in children. Other phonological assessment instruments (CAAP and GFTA-2) have also proved to have substantial discriminative power in samples of children phonological disorders^(10,11).

The development of the INFONO was spurred by the need for standardized phonological assessment tools for the population of Southern Brazil. The instrument was developed as a software package in order to make it more appealing to children and facilitate the administration and calculation of test results by the speech pathologist. Nevertheless, the present study had some limitations, such as the absence of separate reliability coefficients for each normative subgroup (age group, gender, school type). Reliability coefficients can vary depending on the skill level of the normative population, with younger age groups often yielding lower reliability values⁽¹⁴⁾.

In summary, this investigation of the psychometric properties of the INFONO was able to demonstrate the validity and reliability of its administration and scoring, supporting its use as a complementary diagnostic tool. In the absence of a gold-standard method for phonological assessment in Brazil, in the form of another instrument which has undergone validity and reliability testing, no further comparative analyses could be undertaken. However, additional studies should be conducted to seek further evidence of the validity and reliability of the INFONO. Despite these limitations, this study has important implications for the assessment and diagnosis of phonological disorders in southern Brazil, as it presents the INFONO as the first standardized phonological instrument for which evidence of validity and reliability has been reported.

CONCLUSION

The present study provided evidence of the validity (internal consistency and criterion validity - discriminant analysis) and reliability (internal consistency, and intra-/interrater reliability) of the INFONO. These analyses demonstrated that the items in the instrument are reliable and produce consistent scores even when rated by different examiners or by a single examiner at different points in time. These findings may be attributable to our adherence to standard procedures in the administration, recording and scoring of the instrument, and to the experience of the examiners, which would suggest that examiners must be thoroughly familiar with the instrument before they administer it themselves.

These results demonstrate the diagnostic reliability of the normative data provided for the INFONO. Overall, the present findings support the use of this instrument as a valid and reliable tool to evaluate typical and atypical phonological acquisition in both clinical practice and research in speech pathology.

ACKNOWLEDGEMENTS

We would like to thank the following research assistants: Jonatas Corteze, from the Department of Computer Science, for developing different versions of the INFONO to be used at different stages of the project; Matheus Tanuri from the Department of Industrial Design for developing the images in the INFONO; and Matheus Moreira, for developing our logo. We would also like to thank the Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS) for their support to this project, in the form of an Edital PqG grant (number 11/2184-8).

REFERENCES

1. Namasivayam AK, Pukonen M, Goshulak D, Yu VY, Kadis DS, Kroll R, et al. Relationship between speech motor control and speech intelligibility in children with speech sound disorders. *J Commun Disord.* 2013;46(3):264-80. <http://dx.doi.org/10.1016/j.jcomdis.2013.02.003>. PMID:23628222.
2. Strand EA, McCauley RJ, Weigand SD, Stoeckel RE, Baas BS. A motor speech assessment for children with severe speech disorders: reliability and validity evidence. *J Speech Lang Hear Res.* 2013;56(2):505-20. [http://dx.doi.org/10.1044/1092-4388\(2012/12-0094\)](http://dx.doi.org/10.1044/1092-4388(2012/12-0094)). PMID:23275421.
3. Grigos MI, Moss A, Lu Y. Oral articulatory control in childhood apraxia of speech. *J Speech Lang Hear Res.* 2015;58(4):1103-18. http://dx.doi.org/10.1044/2015_JSLHR-S-13-0221. PMID:25951237.
4. Sugden E, Baker E, Munro N, Williams AL. Involvement of parents in intervention for childhood speech sound disorders: a review of the evidence. *Int J Lang Commun Disord.* 2016;51(6):597-625. <http://dx.doi.org/10.1111/1460-6984.12247>. PMID:27017993.
5. Mürsepp I, Aibast H, Gapeyeva H, Pääsuke M. Motor skills, haptic perception and social abilities in children with mild speech disorders. *Brain Dev.* 2012;34(2):128-32. <http://dx.doi.org/10.1016/j.braindev.2011.02.002>. PMID:21376491.
6. Lousada M, Jesus LMT, Hall A, Joffe V. Intelligibility as a clinical outcome measure following intervention with children with phonologically based speech-sound disorders. *Int J Lang Commun Disord.* 2014;49(5):584-601. <http://dx.doi.org/10.1111/1460-6984.12095>. PMID:24861159.
7. Barrozo TF, Pagan-Neves LO, Vilela N, Carvalho RM, Wertzner HF. The influence of (central) auditory processing disorder in speech sound disorders. *Braz J Otorhinolaryngol.* 2016;82(1):56-64. <http://dx.doi.org/10.1016/j.bjorl.2015.01.008>. PMID:26612604.
8. Vick JC, Campbell TF, Shriberg LD, Green JR, Abdi H, Rusiewicz HL, et al. Distinct developmental profiles in typical speech acquisition. *J Neurophysiol.* 2012;107(10):2885-900. <http://dx.doi.org/10.1152/jn.00337.2010>. PMID:22357794.
9. Farquharson K. Language or motor: reviewing categorical etiologies of speech sound disorders. *Front Psychol.* 2015;6:1708. <http://dx.doi.org/10.3389/fpsyg.2015.01708>. PMID: 26583005.
10. Goldman R, Fristoe M. Goldman-Fristoe Test of Articulation 2 - Test manual. Circle Pines, MN: American Guidance Services Inc; 2000.
11. Secord WA, Donohue JS. CAAP - Clinical Assessment of Articulation and Phonology – Supplemental Examiner’s Manual. USA: Super Duper® Publications; 2002.
12. Yavas M, Hernandez CLM, Lamprecht RR. Avaliação Fonológica da Criança. Porto Alegre: Artes Médicas; 2002.
13. Wertzner HF. Fonologia (Parte A). In: Andrade CRF, Befi-Lopes DM, Fernandes FDM, Wertzner HF, editors. Teste de linguagem infantil nas áreas de fonologia, vocabulário, fluência e pragmática. Carapicuíba: Pró-Fono; 2004. Vol. 2, 98 p.
14. Kirk C, Vigeland L. A psychometric review of norm-referenced tests used to assess phonological error patterns. *Lang Speech Hear Serv Sch.* 2014;45(4):365-77. http://dx.doi.org/10.1044/2014_LSHSS-13-0053. PMID:25091265.

15. McLeod S, Verdon S. A review of 30 speech assessments in 19 languages other than English. *Am J Speech Lang Pathol.* 2014;23(4):708-23. http://dx.doi.org/10.1044/2014_AJSLP-13-0066. PMID:24700105.
16. Cunha VLO, Capellini AS. Construção e validação de instrumento de avaliação da compreensão de leitura para escolares do terceiro ao quinto ano do Ensino Fundamental. *CoDAS.* 2014;26(1):28-37. PMID:24714856.
17. Flipsen P Jr, Ogiela DA. Psychometric characteristics of single-word tests of childrens's speech sound production. *Lang Speech Hear Serv Sch.* 2015;46(2):166-78. http://dx.doi.org/10.1044/2015_LSHSS-14-0055. PMID:25654693.
18. Friberg JC. Considerations for test selection: How do validity and reliability impact diagnostic decisions? *Child Lang Teach Ther.* 2010;26(1):77-92. <http://dx.doi.org/10.1177/0265659009349972>.
19. McCauley RJ, Swisher L. Use and misuse of norm-referenced tests in clinical assessment: a hypothetical case. *J Speech Hear Disord.* 1984;49(4):338-48. <http://dx.doi.org/10.1044/jshd.4904.338>. PMID:6389982.
20. Fonseca RP, Parente MAMP, Côté H, Ska B, Joannette Y. Apresentando um instrumento de avaliação da comunicação à Fonoaudiologia Brasileira: bateria MAC. *Pro Fono.* 2008;20(4):285-91. <http://dx.doi.org/10.1590/S0104-56872008000400014>. PMID:19142474.
21. Pagliarini KC, Ortiz KZ, Parente MAMP, Arteché A, Joannette Y, Nespoulous JL, et al. Montreal-Toulouse Language Assessment Battery for aphasia: Validity and reliability evidence. *NeuroRehabilitation.* 2014;34(3):463-71. PMID:24473247.
22. Silva MK, Ferrante C, Borsari JV, Pereira MMB. Aquisição fonológica do Português Brasileiro em crianças do Rio de Janeiro. *J Soc Bras Fonoaudiol.* 2012;24(3):248-54.
23. Ceron MI, Gubiani MB, Oliveira CR, Keske-Soares M. Factors influencing consonant acquisition in Brazilian Portuguese-Speaking Children. *J Speech Lang Hear Res.* 2017;60:759-771. http://dx.doi.org/10.1044/2016_JSLHR-S-15-0208.
24. Savoldi A, Ceron MI, Keske-Soares M. Quais são as melhores palavras para compor um instrumento de avaliação fonológica? *Audiol Commun Res.* 2013;18(3):194-202. <http://dx.doi.org/10.1590/S2317-64312013000300009>.
25. Ceron MI. Instrumento de Avaliação Fonológica (INFONO): Desenvolvimento e estudos psicométricos. Universidade Federal de Santa Maria, Santa Maria; 2015.
26. Pavez MM, Maggiolo M, Coloma CJ. Test para evaluar procesos de implificación fonológica - TEPROSIF-R. Santiago: Pontificia Universidad Católica de Chile; 2009.
27. Gubiani MB, Keske-Soares M. Evolução fonológica de crianças com desvio fonológico submetidas a diferentes abordagens terapêuticas. *Rev CEFAC.* 2014;16(2):663-71. <http://dx.doi.org/10.1590/S1516-18462012005000071>.
28. Giacchini V, Mota HB, Mezzomo CL. Variáveis relevantes no processo terapêutico para a aquisição do onset complexo na fala de crianças com desvio fonológico. *Rev CEFAC.* 2015;17(1, Supl 1):17-26. <http://dx.doi.org/10.1590/s1516-18462012005000049>.
29. Murphy CFB, Pagan-Neves LO, Wertzner HF, Schochat E. Children with speech sound disorder: comparing a non-linguistic auditory approach with a phonological intervention approach to improve phonological skills. *Front Psychol.* 2015;6:64. <http://dx.doi.org/10.3389/fpsyg.2015.00064>. PMID:25698997.
30. Lousada M, Mendes AP, Valente AR, Hall A. Standardization of a phonetic-phonological test for European-Portuguese children. *Folia Phoniatr Logop.* 2012;64(3):151-6. <http://dx.doi.org/10.1159/000264712>. PMID:22965068.

Author contributions

MIC was responsible for study conception, data collection, analysis and interpretation; MBG was responsible for data collection and reviewing the manuscript; CRO contributed to data interpretation and manuscript review; MKS was responsible for study design and for the approval of the final version of the article sent for publication.