Running head: ordinal vs. VAS ratings in resonance disorders

Perceptual evaluation of hypernasality, audible nasal airflow and speech understandability using ordinal and visual analogue scaling and their relation with nasalance scores

Kim Bettens[a], Laura Bruneel[a] , Youri Maryn[b,c,d], Marc De Bodt[a,e], Anke Luyten[a], and Kristiane M. Van Lierde[a,f]

[a]Department of Speech, Language and Hearing Sciences, Ghent University, Ghent, Belgium
[b]Department of Otorhinolaryngology, Speech-Language Pathology, Sint-Augustinus General Hospital, Wilrijk, Antwerp, Belgium
[c]Department of Speech-Language Therapy and Audiology, Faculty of Education, Health and Social Work, University College Ghent, Ghent, Belgium
[d]Faculty of Medicine and Health Sciences, University of Antwerp, Antwerp, Belgium
[e]Rehabilitation Centre for Communication Disorders, Antwerp University Hospital, Wilrijk, Belgium
[f]Department of Speech-Language Therapy and Audiology, University of Pretoria, Pretoria, South-Africa

Correspondence according this article should be addressed to Kim Bettens, Department of Speech, Language and Hearing Sciences, Ghent University, De Pintelaan 185, 2P1, BE-9000 Gent, Belgium. E-mail: Kim.Bettens@Ugent.be, Telephone: +32 9 332 94 26, Fax: +32 9 332 54 36.

Abstract

Purpose

Perceptual assessments remain the most commonly utilized procedure to diagnose and evaluate resonance disorders. However, the discussion continues about which rating scale has to be applied. Therefore, this study aimed to compare the reliability and validity of ordinal and visual analogue scales to rate hypernasality, audible nasal airflow and speech understandability.

Methods

Four experienced speech-language pathologists rated 35 speech samples of children with a range of hypernasality, audible nasal airflow and speech understandability, using an ordinal scale and a visual analogue scale. Intraclass correlations coefficients determined intra- and inter-rater reliability. The model of best fit was determined by plotting both rating scales against each other. A Pearson correlation coefficient verified the relationship between both rating scales and nasalance scores determined by a Nasometer.

Results

Good intra- and inter-rater reliability was found for both rating scales. A multiple regression analysis revealed a curvilinear relationship between both rating scales, indicating a slight preference to rate all parameters by a visual analogue scale. Comparable correlations with nasalance scores were found.

Conclusions

This study confirms that visual analogue scale ratings form a reliable and valid alternative for ordinal ratings in the perceptual judgments of hypernasality, audible nasal airflow and speech understandability. A combination of both rating scales may even combine the advantages and eliminate their limitations. However, further research is necessary to verify how this new approach can be implemented in available protocols for clinical practice, audits and research.

Introduction

To diagnose resonance disorders, perceptual judgments are still considered the gold standard since no instrumental measurement can transcend the capabilities of a trained ear (Henningsson et al., 2008). Hence, perceptual judgments need to be valid and reliable. Recently, initiatives have been taken to standardize perceptual assessments and to explore both validity and reliability of perceptual assessment protocols for resonance disorders in patients with cleft palate. In order to meet the demand for a uniform speech sample, Henningsson et al. (2008) proposed universal parameters that can be applied in several languages to compose a consistent speech sample. However, the reliability and validity of those samples are not yet confirmed. Another speech analysis protocol to judge hypernasality, nasal airflow and articulation, more specifically the Cleft Audit Protocol for Speech-Augmented (CAPS-A), was developed by John, Sell, Sweeney, Harding-Bell, and Williams (2006). Validity, reliability and acceptability of this protocol were proven by several studies (Britton et al., 2014; Chapman et al., 2016; John et al., 2006). All these perceptual assessment protocols include categorical scales with clear description of the different grades to judge hypernasality, audible nasal airflow (ANA) and intelligibility. Moreover, categorical scaling was applied in 74% of the studies that included a perceptual assessment of cleft palate speech, as reported in the critical review of A. Lohmander and Olsson (2004).

However, the discussion continues about the type of rating scale that has to be applied to judge resonance reliably (Baylis, Chapman, Whitehill, & Group, 2015; Brancamp, Lewis, & Watterson, 2010; Whitehill, Lee, & Chun, 2002; Zraick & Liss, 2000). This discussion is not only restricted to the perception of resonance, but also occurs in the domain of other speech dimensions like voice (Michael P. Karnell et al., 2007; Wuyts, De Bodt, & Van de Heyning, 1999; Yiu & Ng, 2004) and stuttering (Schiavetti, Sacco, Metz, & Sitler, 1983). The discussion is related to the mental processing behind the perception of resonance (Stevens, 1975). Stevens (1975) differentiates two kinds of perceptual continua: metathetic and prothetic. Metathetic dimensions involve qualitative perceptual changes,

while prothetic dimensions involve quantitative perceptual changes. As a result, metathetic stimuli are more concerned with 'what kind' (e.g., pitch), whereas prothetic stimuli are concerned with 'how much' (e.g., loudness) (Roeckelein, 1998). Due to this difference in mental processing, Stevens (1975) postulated that different perceptual phenomena should be rated by different scales. More specifically, he proposed to rate metathetic stimuli by partition measures, whereas prothetic stimuli should be rated by using magnitude measures. Examples of partition measures are equal appearing interval (EAI) and ordinal scales, in which the listener chooses between a finite set of numbers (0 to N) or categories to rate a specific stimulus, representing the quality of the stimulus. Magnitude measures, on the other hand, represent quantitative measures by assigning numbers to stimuli in proportion to their magnitude, for example direct magnitude estimation (DME) and visual analogue scales (VAS) (Yiu & Ng, 2004).

For judgments based on EAI scales, whole numbers are used to divide the scale in equal intervals. The endpoints are fixed in which odd-numbered scales with 5 to 7 points are usually applied to rate resonance disorders in clinical practice. When adjectives or definitions are added to the different numbers, intervals are not equally appearing anymore, but are defined by the descriptions, resulting in an ordinal scale. Although both scales are fundamentally categorical (Wuyts et al., 1999), Castick, Knight, and Sell (2017) cautioned that researchers should not use the terms 'equal appearing interval' and 'ordinal' interchangeably.

Categorical (or partition) scales are widely used to rate resonance disorders as they are accessible, easy to use and interpret and it is easy to compare the results of different listeners or patients (Baylis et al., 2015). Nevertheless, the validity of this type of rating scales is questioned as several psychophysical experiments denoted that listeners divide the lower end of the scale into smaller intervals (Stevens, 1975). This results in a systematic bias towards the lower part of the continuum (Brancamp et al., 2010). Additionally, only categorical data are provided which limits the subsequent statistical analysis (Brancamp et al., 2010). Hence, the use of magnitude measures scales,

such as DME and VAS, was explored to rate resonance. DME can be applied with a modulus, or standard for comparison (DME-M), or without a modulus (DME-WM). During DME-M, a specific value is assigned to a standard speech sample (i.e. the modulus) after which the listener rates all speech samples relative to the magnitude of the modulus. For example, if the value of the modulus was set at 100 and a subsequent speech sample is judged to be twice as nasal as the modulus, a value of 200 will be given to that sample. When a DME-WM is used, the listener has to assign a value to the first sample, after which all other samples are compared with this first. As this is a ratio scale, no systematic bias associated with categorical scales occurs. Moreover, there are more options regarding statistical analyses (Brancamp et al., 2010). Nevertheless, the procedure is complex because it requires more explanation and training of the listeners, a more complicated presentation of the stimulus and complex statistical analysis (Baylis et al., 2015). Therefore, implementation of this rating procedure in a clinical setting seems difficult. Hence, VAS was explored by Baylis et al. (2015) as an alternative magnitude measures scale to judge resonance disorders since VAS and DME ratings of hypernasality seem to be strongly correlated (Cheng, 2006). A 100mm bar is presented to the listeners who have to place a mark on the bar going from 0 (normal) to 100 (most severely disturbed). This results in a continuous level of measurement suggesting an easier implementation in daily clinical practice because of its ease of use and the more convenient analysis (Baylis et al., 2015).

To determine whether a perceptual phenomenon is prothetic or metathetic, Stevens (1975) postulated to explore the relationship between partition and magnitude scale ratings based on judgments of the same speech sample. If the ratings are related to each other in a linear way, the rated phenomenon can be considered to be metathetic. If a quadratic or curvilinear relationship is found, the rated phenomenon would be prothetic. In the literature, most of the studies reported a curvilinear relationship between partition and magnitude measure scale ratings of hypernasality, suggesting that hypernasality is rather a prothetic phenomenon that can be rated more validly by using magnitude measures scales such as DME or VAS (Baylis et al., 2015; Baylis, Munson, & Moller, 2011; Whitehill et al., 2002; Zraick & Liss, 2000, see supplementary material). Brancamp et al. (2010), on the other hand,

reported no statistically significant differences between ratings of hypernasality based on EAI and DME scales based on the judgments of one rater and found a linear relationship between EAI and DME ratings. Furthermore, Castick et al. (2017) found only a slightly stronger curvilinear than linear relationship based on ordinal scales and VAS, and this for six dimensions of cleft speech.

As mentioned above, ordinal and EAI scales, although both partition scales, are based on different principles. Therefore, caution is advised to compare the results between studies using different kind of partition scales. Although most of the internationally accepted protocols for perceptual judgments of cleft speech parameters include ordinal scales (Chapman et al., 2016; Henningsson et al., 2008; John et al., 2006; Anette Lohmander, Lundeborg, & Persson, 2017), only two studies (Baylis et al., 2015; Castick et al., 2017) to date have included ordinal scales to explore their relationship with magnitude measures. Both studies reported a curvilinear relationship between ordinal scaling and VAS. However, Castick et al. (2017) only found a slightly stronger curvilinear than linear relationship. As a result, the conclusions of both studies differed. While Baylis et al. (2015) favor the use of VAS, Castick et al. (2017) stated that both scales are appropriate for measuring cleft speech parameters. A possible explanation for this discrepancy is the difference in applied method to determine the model of best fit. In the study by Baylis et al. (2015) all individual ratings of the five listeners were included to determine the model of best fit, whereas Castick et al. (2017) used mean values. However, no rationale was provided by the authors for their choice.

Regarding intra- and inter-rater reliability, most authors reported some higher reliability scores for the magnitude measures scales compared to the partition measures scales when rating hypernasality or ANA (Baylis et al., 2015; Baylis et al., 2011; Brancamp et al., 2010; Whitehill et al., 2002; Zraick & Liss, 2000), although the reliability of the EAI and ordinal scales mostly stays acceptable. Hence, no consensus is yet reached about which scale should be used to judge cleft speech parameters. Despite the growing evidence that magnitude measures scales are more appropriate than partition measures scales to rate hypernasality and ANA, partition measures scales are still widely used, even in

internationally accepted protocols as stated above. A possible explanation may be the limited research regarding VAS as an alternative for the more complex DME ratings. If more evidence is provided in favor of VAS, researchers and clinicians may be more encouraged to revise current assessment protocols.

Furthermore, it is interesting to explore the correlation between perceptual judgments based on magnitude measures scales and instrumental assessments. As Brancamp et al. (2010) stated, nasalance scores are mostly compared with partition measures ratings. However, if hypernasality is rather a prothetic than a metathetic phenomenon, partition measures ratings will be less valid to rate this perception which may result in lower correlations between perceptual and instrumental measurements. A common instrument to provide indirect measures of hypernasality is the Nasometer, originally developed by Fletcher and Bishop (1973) and manufactured by KayPentax (NJ, Lincoln Park). The Nasometer measures the amount of nasalance by capturing the oral and nasal signal using two microphones on a sound separation plate which is placed under the nose of the patient. After bandpass filtering, the nasal signal is divided by the total signal of oral and nasal energy and multiplied by 100 to receive the nasalance score in percentage. Reported correlation coefficients between partition measures ratings (i.e. EAI or ordinal scales) of hypernasality and nasalance scores of oral stimuli vary between 0.29 to 0.76, with most authors reporting moderate correlations (Brancamp et al., 2010 – r=0.63; Brunnegard, Lohmander, & van Doorn, 2012 – r=0.49-0.76; Dalston, Neiman, & Gonzalezlanda, 1993 – r=0.73; Lewis, Watterson, & Houghton, 2003 – r=0.29-0.57; Sweeney & Sell, 2008 – r=0.74; Watterson, McFarlane, & Wright, 1993 – r=0.49). Only three studies compared magnitude measures scale (i.e. VAS) scores with nasalance scores. In each of these studies, low to moderate correlations were reported (Bettens et al., 2016 – r=0.63; Brancamp et al., 2010 – r=0.59; Keuning, Wieneke, van Wijngaarden, & Dejonckere, 2002 – r=0.36-0.60). Nevertheless, no comparison has yet been made between the correlations of nasalance scores and partition measures ratings versus nasalance scores and magnitude measures ratings.

Although partition measures scales are most commonly used to judge resonance disorders in clinical practice and research, this review of the literature shows that this type of rating scales is currently being questioned. However, additional evidence is needed before this strongly embedded and widely accepted method would be replaced. Therefore, the aims of the present study were: (1) to determine intra- and inter-rater reliability of ordinal and VAS ratings of hypernasality, ANA and speech understandability based on judgments of experienced speech-language pathologists (SLP); (2) to determine the model of best fit between ordinal and VAS ratings of hypernasality, ANA and speech understandability in order to contribute to the question if those percepts are rather prothetic than metathetic phenomena; (3) to explore the correlation between ordinal and VAS ratings of hypernasality and nasalance scores of an oral and oronasal text obtained by a Nasometer. Based on literature, we hypothesize that the reliability of the VAS ratings will be comparable with the reliability of the ordinal ratings. Furthermore, we hypothesize a curvilinear relationship between the ordinal and VAS ratings. Finally, somewhat higher correlations are expected between the VAS ratings of hypernasality and the nasalance scores compared to the correlation between the ordinal ratings of hypernasality and the nasalance scores.

Method

This study was part of a larger study regarding perceptual and instrumental assessments of resonance disorders in children with cleft palate and children with non-cleft related velopharyngeal insufficiency. It was approved by the institutional review and ethical board of the xx University Hospital (EC/2012/049). One listening experiment, including four experienced SLPs who judged 35 speech samples using both ordinal ratings and VAS, was set up to meet the objectives of two different studies. A first study aimed to explore the validity of a new instrumental assessment procedure to determine hypernasality, more specifically the Nasality Severity Index 2.0 (NSI 2.0) (Bettens et al., 2016). Therefore, the perceptual judgments of hypernasality based on VAS were used to determine the correlation with the NSI 2.0 scores. While the focus of this first study was on the instrumental

assessment of resonance, the current study focused on the perceptual judgment of resonance. Hence, the perceptual judgments based on both VAS and ordinal ratings were applied in the current study to compare their reliability and validity. Because the methodological procedure was already described in the previous study, only a summary of the method is provided below. For more details, please refer to the previous study (Bettens et al., 2016).

Listeners. Four SLPs with at least 5 years of experience in rating resonance disorders served as listeners. They all were native speakers of Flemish Dutch and worked with patients with resonance disorders, both in a clinical as well as a research setting. They all had experience in rating resonance disorders by using ordinal rating scales. However, none of them had used VAS to judge resonance disorders before. Nevertheless, three of the four listeners did have experience in using VAS to judge voice disorders.

Speakers. Speech samples were collected from 35 children between 4 and 15 years old (mean age 7.3y, SD 2.67), representing a range of hypernasality and ANA, from absent to severely present. ANA included audible nasal emission as well as nasal turbulence. Some of the children also had articulatory difficulties which influenced speech understandability. All children were referred to the department of speech and language pathology at the xx University Hospital in Belgium with a complaint of hypernasal speech due to a variety of pathologies or during a follow-up period after palatal repair (see Bettens et al. (2016) for more details regarding the pathologies). The inclusion criteria to participate in this study were being a native speaker of Dutch, living in Flanders (the northern part of Belgium), and being able to produce the required speech sample. Children suffering from a cold or congestion at the moment of testing or presenting with hyponasal resonance, a pharyngeal flap, learning disabilities greater than mild, dysarthria or dyspraxia were excluded from the study.

Speech samples. A speech sample based on spontaneous speech was collected from each child. Conversational speech was chosen to provide representative information about the articulation and resonance (Kuehn & Moller, 2000). The first 65 syllables of each sample (i.e. the length of the

smallest available sample) were selected, resulting in speech samples with a similar length in terms of number of syllables. All samples were video-recorded using a Sony HDR-CX280 camera in a quiet room at the clinical department of the xx University Hospital. To limit listener bias related to the child's appearance, all samples were converted to audio samples using audio converter software (Freemake Audio Converter, version 1.1.0.66) at a sampling frequency of 48kHz.

Instrumental assessment. Following the procedure described by Bettens et al. (2016), a Nasometer 6450 model II (Kay Pentax, USA) was used to determine nasalance values of two text passages. The first passage, a so called 'oral' text, exclusively consists of oral speech sounds and is used to detect hypernasality. The second passage, the 'oronasal' text, contains approximately the same percentage of nasal phonemes (11.67%) as in spontaneous Dutch speech (11.63%, Van den Broecke, 1988)). The passages were originally developed by Van de Weijer and Slis (1991) and are available in the appendix.

Rating procedures. Each listener was asked to rate the degree of hypernasality, the frequency of ANA (including audible nasal emission and/or nasal turbulence) and the degree of speech understandability of each speech sample. As described by Bettens et al. (2016), a short training session was given by the first author before starting the listening experiment. During this training session, ratings were first performed using VAS. For each sample, the ends of the 100mm line were labeled as 'normal' (or 'absent') on the left end and 'severely distorted' (or 'frequently noted') on the right end. Second, the same samples were judged using an ordinal scale. Hypernasality was rated based on five categories (0 = absent, 1 = borderline, 2 = mild, 3 = moderate, 4 = severe); the frequency of occurrence of ANA was judged on a three-point scale (0 = absent, 1 = mild, 2 = marked), both based on the definitions and rating system of the CAPS-A (John et al., 2006). Speech understandability was judged on a four-point scale (0 = within normal limits, 1 = mild, 2 = moderate, 3 = severe) based on the definitions provided by Henningsson et al. (2008). None of the example samples were included in the study samples.

After completing the training session, the listening experiment was completed as described by Bettens et al. (2016). Each listener received a standard pair of over-ear headphones (Sennheiser EH150) and the blinded audio samples in a randomized sequence to minimize order effects. Samples were played from a personal computer (Dell Latitude, Microsoft Windows 10) using the Windows Media Player. To verify intra-rater reliability, 26% of the samples (9/35) were repeated. To minimize order effects, two raters were at random assigned to rate the samples using the VAS first, while the other two raters used the ordinal scale first. A short break was inserted after completing 22 speech samples after which the following 22 speech samples were rated using the same rating scale. After completion of all samples, a break of two hours was taken. Subsequently, the same set of samples was presented to the raters in another randomized order while they were asked to rate the samples with the opposite rating scale. In total, the rating task took about one and a half hours per rating condition.

Data analysis. Intraclass correlation coefficients (ICCs) were calculated to determine the intra- and inter-rater reliability of both rating scales for the variables hypernasality, ANA and understandability (Hallgren, 2012). The intra-rater reliability was determined using a two-way mixed model (ICC (3,1)) following the classification of Shrout and Fleiss (1979) and using SPSS software version 22.0 (SPSS Inc., IBM PC version). Inter-rater reliability was determined using a two-way mixed model (ICC (3,k)).

To determine the model of best fit between the ordinal and VAS ratings, the results of both rating scales were plotted against each other in accordance with the procedure provided by Stevens (1975), using the individual ratings of all listeners. The choice of using the individual ratings was based on the possible influence of mean ratings on the fundamental relationship between two sets of ratings (Brancamp et al., 2010). More specifically, mean ratings can result in a shift toward the center of the scale because the extreme scale points are often not chosen by all listeners. Additionally, the higher end of a scale becomes underrepresented relative to the low end because listeners agree more often in normal samples. These effects may give the statistical impression that the relationship between

both rating scales is not linear, resulting in the conclusion that magnitude measures like VAS are more appropriate to rate resonance than partition measures like ordinal scales. To determine the model of best fit statistically, a multiple regression analysis was performed in which linear and quadratic terms were subsequently entered into the model. A statistically significant change of R^2 after the addition of the quadratic term indicates a curvilinear relationship.

Finally, a Pearson correlation coefficient was used to verify the relationship between the degree of hypernasality based on both ratings scales and the nasalance scores of the oral and oronasal texts. Additionally, a Fisher's r-to-Z transformation was used to compare the correlation coefficients of both rating scales.

Results

Reliability. Table 1 shows the ICCs and 95% confidence intervals (95% CI) for the *intra-rater reliability* for both rating scales used to judge hypernasality, frequency of ANA and understandability. Based on the guidelines by Cicchetti (1994), good to excellent agreement was found for judgments based on ordinal scaling for all judged parameters. Regarding VAS, good to excellent agreement was found for the judgments of hypernasality. However, only a fair agreement was found for the judgments of this variable by listener 4. Furthermore, excellent agreement was found for the ANA and understandability judgments. Average-measures ICCs and their 95% CIs for the *inter-rater reliability* are provided in Table 2. For both rating scales excellent levels of agreement were found for the ratings of hypernasality and understandability, and good levels of agreement were found for the ratings of ANA.

Model of best fit. In accordance with the procedure described by Stevens (1975), Figures 1 to 3 provide the graphic reproduction of the VAS rating scores relative to the ordinal rating scores based on judgments of the degree of hypernasality, ANA and understandability. A linear as well as curvilinear relationship are graphically presented. Slightly higher, but significantly different R^2 values for the line of best fit indicate a slight advantage of the curvilinear model over the linear model for the judgments

of hypernasality based on individual ratings (linear model: R^2=0.579, curvilinear model: R^2=0.595, p=0.021, Figure 1) and ANA (linear model: R^2=0.504, curvilinear model: R^2=0.523, p=0.023, Figure 2). Non-significantly different R^2 values for the line of best fit were found for the judgments of understandability (linear model: R^2=0.785, curvilinear model: R^2=0.585,p=0.965, Figure 3).

Hypernasality and nasalance. A significant, moderate correlation (r=0.522, p=0.001) was found between the mean ordinal ratings of the degree of perceived hypernasality and nasalance scores of the oral text. A somewhat higher, although still moderate correlation (r=0.629, p<0.001) was found between the mean VAS ratings and the nasalance scores of the oral text. Nevertheless, no statistically significant difference was detected between both correlation coefficients (z=-0.64, p=0.520). Regarding the nasalance scores of the oronasal text, a small, but significant correlation was found with the mean ordinal ratings of the degree of perceived hypernasality (r=0.370, p=0.031) and a moderate, significant correlation was found with the mean VAS ratings of the degree of perceived hypernasality (r=0.514, p=0.002). Nevertheless, no statistically significant difference was detected between both correlation coefficients (z=-0.71, p=0.478).

Discussion

This study aimed to provide additional evidence to support the hypothesis that hypernasality, ANA and speech understandability are perceived as prothetic instead of metathetic phenomena and therefore can be rated more reliably using magnitude scales such as VAS instead of partition scales such as ordinal scales. It is the third study (Baylis et al., 2015; Castick et al., 2017) that compared perceptual judgments of hypernasality and ANA based on ordinal scales and VAS. However, it is the first that compared the correlations between nasalance scores and perceptual judgments based on different rating scales. Furthermore, it is the first study about this topic that was conducted in a language other than English.

A regression analysis based on the individual rating points of all listeners showed that a curvilinear model best fits the relationship between the perceptual judgments based on an ordinal

scale and VAS for both hypernasality and ANA. However, although significant, the contribution of the curvilinear model to explain the amount of variance was very limited. These results are in line with the results reported in previous studies (Baylis et al., 2015; Castick et al., 2017), although the difference between the contribution of the linear and curvilinear model was more clear in the study by Baylis et al. (2015). Regarding the parameter 'understandability', no significant contribution of the curvilinear model was found, which is comparable with the findings reported by Castick et al. (2017). In accordance with the study by Castick et al. (2017), these findings provide additional evidence that hypernasality, ANA and understandability can be rated equivalently by using both VAS and ordinal scales.

Despite the less intensive listening training of the raters in the present study, overall good to excellent levels of intra- and inter-rater reliability were found for both the ordinal and VAS ratings, which is also comparable with the results of Baylis et al. (2015) and Castick et al. (2017). Nevertheless, the intra-rater reliability of some listeners was rather low on some parameters which strengthens the need for the development and use of a comprehensive training program and clear definitions to rate resonance and nasal airflow disorders (Chapman et al., 2016; Sell et al., 2009).

Despite the comparable results between the current study and the studies by Baylis et al. (2015) and Castick et al. (2017), some methodological differences need to be addressed. First, as mentioned above, no standard listening training was provided to the raters which may have influenced the reliability of the ratings. Second, the content and length of the speech samples differed. In the current study, only conversational speech was used for the perceptual ratings whereas a combination of conversational speech, sentence repetition and counting was used in the other two studies. Although samples of conversational speech may provide representative information about the articulation and resonance (Kuehn & Moller, 2000), sentence repetition and counting allows for control of the phonetic content (Sell, 2005). Because eliciting conversational speech in young children is not always easy, the more restricted and varying samples in the current study could have influenced the

ratings of the listeners. Additionally, only audio samples were provided to the listeners in the current study and the study by Castick et al. (2017), whereas Baylis et al. (2015) used video samples. This may explain that the results of the current study, to a certain extent, are more comparable with the results of Castick et al. (2017). However, in line with Baylis et al. (2015) but in contrast to Castick et al. (2017), analyses were based on individual instead of mean ratings. Given that mean ratings may favor the statistical impression of a non-linear relationship between two rating scales (Brancamp et al., 2010) as described in the introduction, attention has to be paid to the applied method when interpreting and comparing the results of a specific study. Nevertheless, comparable results were still found with the study by Castick et al. (2017).

Regarding the relationship between hypernasality and nasalance scores of the oral and oronasal texts, slightly higher correlations were found for the judgments of hypernasality using VAS (r=0.63 and r=0.51, respectively) compared to the ordinal ratings (r=0.52 and r=0.37, respectively). Nevertheless, these differences were not statistically significant, suggesting that judgments based on both rating scales comparably correlate with instrumental measurements of hypernasality. For the oral text, the results are comparable with the results reported by Brancamp et al. (2010) who compared judgments of hypernasality by using DME and EAI ratings with the nasalance scores of an oral text. These results indicate that the moderate correlations between perceptual judgments of hypernasality and nasalance scores may not be due to the applied rating scale. As stated by several authors (Karnell, 1995; Sweeney, 2011; Sweeney & Sell, 2008; Watterson, Lewis, & Deutsch, 1998), the presence of ANA, and in particular nasal turbulence, may influence the relationship between perceptual and instrumental measurements. More specifically, the Nasometer cannot discriminate between acoustic energy from nasal resonance and energy from aerodynamic phenomena (such as audible nasal emission and turbulence), nasalance scores may be increased in children with audible nasal airflow problems which can cause inconsistency with listeners' judgments (Watterson et al., 1998). Because several children in this study presented with ANA, this may have had an influence on the correlations between the perceptual ratings of hypernasality and the nasalance scores. Another possible

explanation for the moderate correlations is the use of different stimuli because the phonetic content of a speech sample may influence the perception of hypernasality (Hutters & Henningsson, 2004). More specifically, perceptual judgments were based on spontaneous speech and nasometric values were based on the repetition or reading of an oral text in the current study. Nevertheless, Brunnegard et al. (2012) reported comparable, high correlations between different stimuli (i.e. spontaneous speech vs. nasalance score of oral sentences, r=0.74) and similar stimuli (i.e. oral sentences vs. nasalance scores of oral sentences, r=0.72) using a 5-point ordinal scale.

Taking into account the model of best fit, the reliability and the correlation with the instrumental assessment, we can conclude that hypernasality, ANA and speech understandability can be rated equivalently by using VAS and ordinal scaling. As mentioned by other authors (Baylis et al., 2015; Castick et al., 2017; Cheng, 2006), the use of VAS has several advantages. First, VAS provides more rating options resulting in a higher degree of freedom for the listener (Wuyts et al., 1999). Moreover, the use of VAS creates more opportunities for the statistical analysis of the data because it provides continuous data which allow for parametric analyses, including the possibility of gaining a higher power, which favors this rating scale for research purposes. Nevertheless, reporting that a child has a score of, for example, 38mm on the hypernasality scale may be more difficult to interpret for a parent or other clinician and therefore may be less convenient in clinical practice. Following this, Castick et al. (2017) proposed a combination of both types of rating scales, such as a graphic rating scale (Scott & Huskisson, 1976) which includes descriptors that are spread out along a horizontal line or the incorporation of a color coding system. Another available scale that combines a categorical (C) and ratio (R) scale is the Borg CR10 or the Borg CR100 scale (Borg & Borg, 2001). This scale links verbal anchors to a ratio scale and provides the possibility to rate a specific sample even higher than the fixed maximum of the scale which encourages the listener to use the full length of the scale. However, further research will be needed to define the anchors with attention to correct interpretation and preciseness before this scale can be implemented in daily practice.

This study confirms that VAS ratings form a reliable and valid alternative to ordinal ratings in the perceptual judgments of hypernasality, audible nasal airflow and speech understandability. A combination of both VAS and ordinal scale ratings may even combine the advantages of both rating scales and eliminate their limitations. However, further research is necessary to verify how this new approach can be implemented in available protocols for clinical practice, audits and research. Additionally, this study showed that attention has to be paid to methodological differences regarding the applied speech samples, exact type of rating scales and statistical analyses when the results of similar studies regarding the use of different rating scales are compared.

Appendix

Oronasal Text

Papa en Marloes staan op het station.

Ze wachten op de trein.

Eerst hebben ze een kaartje gekocht.

Er stond een hele lange rij, dus dat duurde wel even.

Nu wachten ze tot de trein eraan komt.

Het is al vijf over drie, dus het duurt nog vier minuten.

Er staan nog veel meer mensen te wachten.

Marloes kijkt naar links, in de verte ziet ze de trein al aankomen.

Oral Text

Het is zaterdag.

Els heeft vrij.

Ze loopt door de stad.

Het is prachtig weer, de lucht is blauw.

Op straat ziet ze Bart op de fiets.

Hij wacht voor het rode licht.

Als Bart haar ziet, zwaait hij.

Els loopt weer verder.

Bij de bakker koopt ze brood, bij de slager koopt ze vlees.

Als het vijf uur is, gaat ze terug, zodat ze op tijd weer thuis is.

References

Baylis, A. L., Chapman, K., Whitehill, T. L., & Group, T. A. S. (2015). Validity and reliability of visual analog scaling for assessment of hypernasality and audible nasal emission in children with repaired cleft palate. *The Cleft Palate-Craniofacial Journal, 52*(3), in press.

Baylis, A. L., Munson, B., & Moller, K. T. (2011). Perceptions of audible nasal emission in speakers with cleft palate: a comparative study of listener judgments. *The Cleft Palate-Craniofacial Journal, 48*(4), 399-411. doi:10.1597/09-201

Bettens, K., De Bodt, M., Maryn, Y., Luyten, A., Wuyts, F. L., & Van Lierde, K. M. (2016). The relationship between the Nasality Severity Index 2.0 and perceptual judgments of hypernasality. *Journal of Communication Disorders, 62*, 67-81. doi:10.1016/j.jcomdis.2016.05.011

Borg, G., & Borg, E. (2001). A new generation of scaling methods: Level-anchored ratio scaling. *Psychologica, 28*(1), 15-45.

Brancamp, T. U., Lewis, K. E., & Watterson, T. (2010). The Relationship Between Nasalance Scores and Nasality Ratings Obtained With Equal Appearing Interval and Direct Magnitude Estimation Scaling Methods. *The Cleft Palate-Craniofacial Journal, 47*(6), 631-637.

Britton, L., Albery, L., Bowden, M., Harding-Bell, A., Phippen, G., & Sell, D. (2014). A cross-sectional cohort study of speech in five-year-olds with cleft palate +/- lip to support development of national audit standards: benchmarking speech standards in the United Kingdom. *The Cleft Palate-Craniofacial Journal, 51*(4), 431-451. doi:10.1597/13-121

Brunnegard, K., Lohmander, A., & van Doorn, J. (2012). Comparison between perceptual assessments of nasality and nasalance scores. *International Journal of Language & Communication Disorders, 47*(5), 556-566. doi:10.1111/j.1460-6984.2012.00165.x

Castick, S., Knight, R.-A., & Sell, D. (2017). Perceptual Judgments of Resonance, Nasal Airflow, Understandability, and Acceptability in Speakers With Cleft Palate: Ordinal Versus Visual Analogue Scaling. *The Cleft Palate-Craniofacial Journal, 54*(1), 19-31.

Chapman, K. L., Baylis, A., Trost-Cardamone, J., Cordero, K. N., Dixon, A., Dobbelsteyn, C., . . . Sell, D. (2016). The Americleft Speech Project: A Training and Reliability Study. *The Cleft Palate-Craniofacial Journal, 53*(1), 93-108. doi:10.1597/14-027

Cheng, T. H. D. (2006). *Direct magnitude estimation versus visual analogue scaling in the perceptual rating of hypernasality.* (bachelor), University of Hong Kong,

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological assessment, 6*(4), 284.

Dalston, R. M., Neiman, G. S., & Gonzalezlanda, G. (1993). Nasometric sensitivity and specificity - a cross-dialect and cross-culture study. *The Cleft Palate-Craniofacial Journal, 30*(3), 285-291. doi:10.1597/1545-1569(1993)030<0285:nsasac>2.3.co;2

Fletcher, S. G., & Bishop, M. E. (1973). Measurement of Nasality with Tonar. *Cleft Palate Journal, 10*, 610-621.

Freemake Audio Converter. http://www.freemake.com/nl/free_audio_converter/.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in Quantitative Methods for Psychology, 8*(1), 23-.

Henningsson, G., Kuehn, D. P., Sell, D., Sweeney, T., Trost-Cardamone, J. E., & Whitehill, T. L. (2008). Universal parameters for reporting speech outcomes in individuals with cleft palate. *The Cleft Palate-Craniofacial Journal, 45*(1), 1-17.

Hutters, B., & Hanningsson, G. (2004). Speech outcome following treatment in cross-linguistic cleft palate studies: methodological implications. *The Cleft Palate–Craniofacial Journal, 41*, 544–549.

John, A., Sell, D., Sweeney, T., Harding-Bell, A., & Williams, A. (2006). The cleft audit protocol for speech-augmented: A validated and reliable measure for auditing cleft speech. *The Cleft Palate-Craniofacial Journal, 43*(3), 272-288. doi:10.1597/04-141.1

Karnell, M. P. (1995). Nasometric discrimination of hypernasality and turbulent nasal airflow. *The Cleft Palate-Craniofacial Journal, 32*(2), 145-148.

Karnell, M. P., Melton, S. D., Childes, J. M., Coleman, T. C., Dailey, S. A., & Hoffman, H. T. (2007). Reliability of clinician-based (GRBAS and CAPE-V) and patient-based (V-RQOL and IPVI) documentation of voice disorders. *Journal of Voice, 21*(5), 576-590.

Keuning, K., Wieneke, G. H., van Wijngaarden, H. A., & Dejonckere, P. H. (2002). The correlation between nasalance and a differentiated perceptual rating of speech in Dutch patients with velopharyngeal insufficiency. *The Cleft Palate-Craniofacial Journal, 39*(3), 277-284. doi:10.1597/1545-1569(2002)039<0277:tcbnaa>2.0.co;2

Kuehn, D. P., & Moller, K. T. (2000). Speech and language issues in the cleft palate population: the state of the art. *The Cleft Palate-Craniofacial Journal, 37*(4), 348.341-348.333.

Lewis, K. E., Watterson, T. L., & Houghton, S. M. (2003). The influence of listener experience and academic training on ratings of nasality. *Journal of Communication Disorders, 36*(1), 49-58.

Lohmander, A., Lundeborg, I., & Persson, C. (2017). SVANTE–The Swedish Articulation and Nasality Test–Normative data and a minimum standard set for cross-linguistic comparison. *Clinical Linguistics & Phonetics, 31*(2), 137-154.

Lohmander, A., & Olsson, M. (2004). Methodology for perceptual assessment of speech in patients with cleft palate: a critical review of the literature. *The Cleft Palate-Craniofacial Journal, 41*(1), 64-70.

Roeckelein, J. E. (1998). *Dictionary of theories, laws, and concepts in psychology*. Westport: Greenwood Publishing Group.

Schiavetti, N., Sacco, P. R., Metz, D. E., & Sitler, R. W. (1983). Direct magnitude estimation and interval scaling of stuttering severity. *Journal of Speech, Language, and Hearing Research, 26*(4), 568-573.

Scott, J., & Huskisson, E. C. (1976). Graphic representation of pain. *Pain, 2*(2), 175-184.

Sell, D. (2005). Issues in perceptual speech analysis in cleft palate and related disorders: a review. *International Journal of Language and Communication Disorders, 40*(2), 103-121.

Sell, D., John, A., Harding-Bell, A., Sweeney, T., Hegarty, F., & Freeman, J. (2009). Cleft audit protocol for speech (CAPS-A): a comprehensive training package for speech analysis. *International Journal of Language & Communication Disorders, 44*(4), 529-548. doi:10.1080/13682820802196815

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428.

Stevens, S. S. (1975). *Psychophysics: introduction to its perceptual, neural, and social prospects*. New York: Wiley.

Sweeney, T. (2011). Nasality - assessment and intervention. In S. Howard & A. Lohmander (Eds.), *Cleft palate speech: assessment and intervention*. Chichester: John Wiley & Sons Ltd.

Sweeney, T., & Sell, D. (2008). Relationship between perceptual ratings of nasality and nasometry in children/adolescents with cleft palate and/or velopharyngeal dysfunction. *International Journal of Language & Communication Disorders, 43*(3), 265-282. doi:10.1080/13682820701438177

Watterson, T., Lewis, K. E., & Deutsch, C. (1998). Nasalance and nasality in low pressure and high pressure speech. *The Cleft Palate-Craniofacial Journal, 35*(4), 293-298.

Watterson, T., McFarlane, S. C., & Wright, D. S. (1993). The relationship between nasalance and nasality in children with cleft palate. *Journal of Communication Disorders, 26*(1), 13-28.

Whitehill, T. L., Lee, A. S. Y., & Chun, J. C. (2002). Direct magnitude estimation and interval scaling of hypernasality. *Journal of Speech, Language, and Hearing Research, 45*, 80-88.

Wuyts, F. L., De Bodt, M. S., & Van de Heyning, P. H. (1999). Is the reliability of a visual analog scale higher than an ordinal scale? An experiment with the GRBAS scale for the perceptual evaluation of dysphonia. *Journal of Voice, 13*(4), 508-517.

Yiu, E. M. L., & Ng, C.-Y. (2004). Equal appearing interval and visual analogue scaling of perceptual roughness and breathiness. *Clinical Linguistics & Phonetics, 18*(3), 211-229.

Zraick, R. I., & Liss, J. M. (2000). A comparison of equal-appearing interval scaling and direct magnitude estimation of nasal voice quality. *Journal of Speech, Language, and Hearing Research, 43*(4), 979-988.

Tables

Table 1. Intra-rater reliability for perceptual ratings of hypernasality, audible nasal airflow (ANA) and understandability based on spontaneous speech using an ordinal or visual analogue scale (VAS).

		Listener 1		Listener 2		Listener 3		Listener 4	
	Rating scale	**Single-Measures ICC**	**95% CI**	**Single-Measures ICC**	**95% CI**	**Single-Measures ICC**	**95% CI**	**Single-Measures ICC**	**95% CI**
Hypernasality	Ordinal	0.63	0.07-0.90	0.93	0.73-0.98	0.85	0.48-0.96	0.95	0.82-0.99
	VAS	0.60	-0.05-0.89	0.90	0.63-0.98	0.93	0.73-0.98	0.42	-0.34-0.85
ANA	Ordinal	1.00	N.A.	0.74	0.20-0.94	0.64	-0.01-0.91	0.69	0.10-0.92
	VAS	0.94	0.77-0.99	0.96	0.83-0.99	0.92	0.70-0.98	0.93	0.73-0.98
Understandability	Ordinal	0.79	0.25-0.95	0.95	0.82-0.99	0.88	0.57-0.97	0.87	0.53-0.97
	VAS	0.98	0.91-1.00	0.86	0.49-0.97	0.96	0.84-0.99	0.90	0.63-0.98

Table 2. Inter-rater reliability for perceptual ratings of hypernasality, audible nasal airflow (ANA) and understandability based on spontaneous speech using an ordinal or visual analogue scale (VAS).

	Rating scale	**Average-Measures ICC**	**95% CI**	**Level of agreement***
Hypernasality	Ordinal	0.82	0.68-0.90	Excellent
	VAS	0.87	0.78-0.93	Excellent
ANA	Ordinal	0.71	0.51-0.84	Good
	VAS	0.74	0.56-0.86	Good
Understandability	Ordinal	0.95	0.90-0.97	Excellent
	VAS	0.92	0.87-0.96	Excellent

*based on Cicchetti (1994): excellent: 0.75-1.00, good: 0.60-0.74, fair: 0.40-0.59, poor: <0.40

Figures

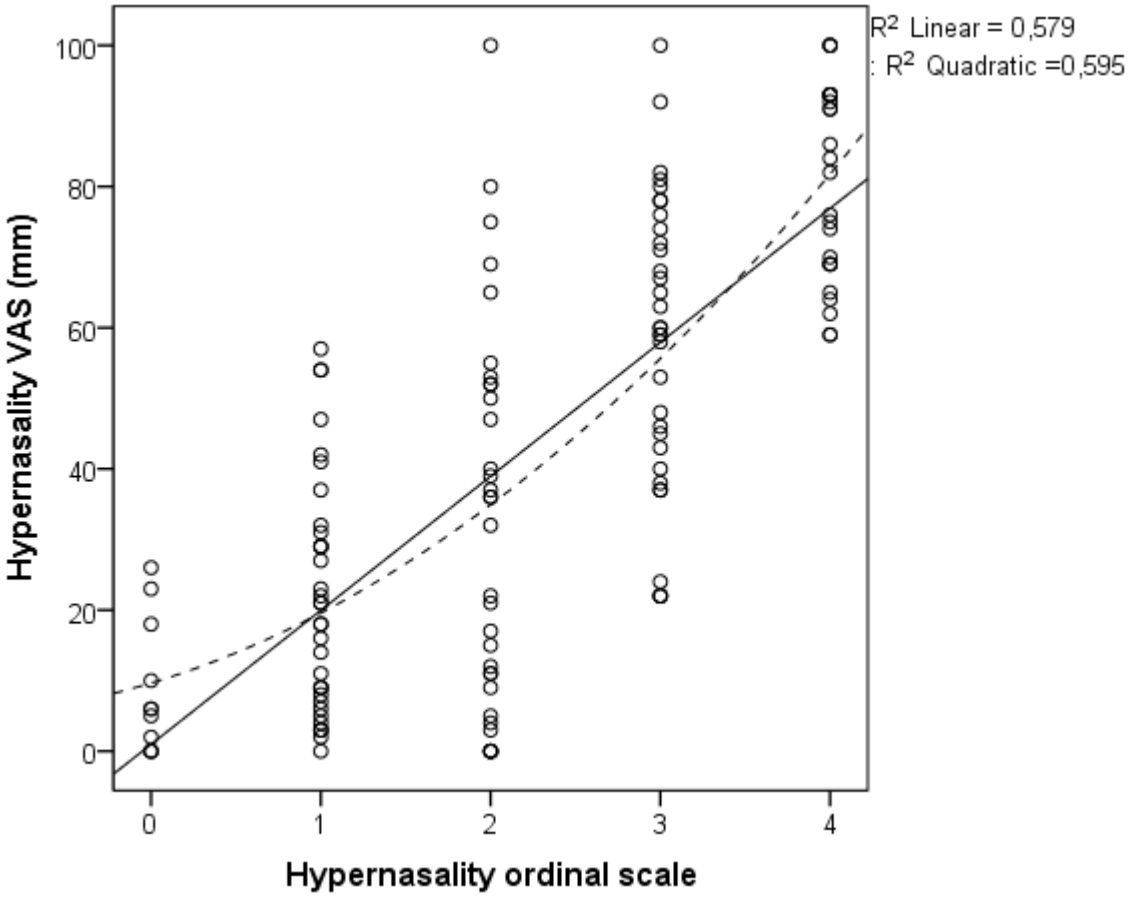

Figure 1. Individual hypernasality ratings based on visual analogue scale (VAS) plotted against individual hypernasality ratings based on an ordinal scale. Linear ($y=0.987 + 18.983x$) and curvilinear ($y=9.666 + 7.221x + 2.702x^2$) relationships are provided.

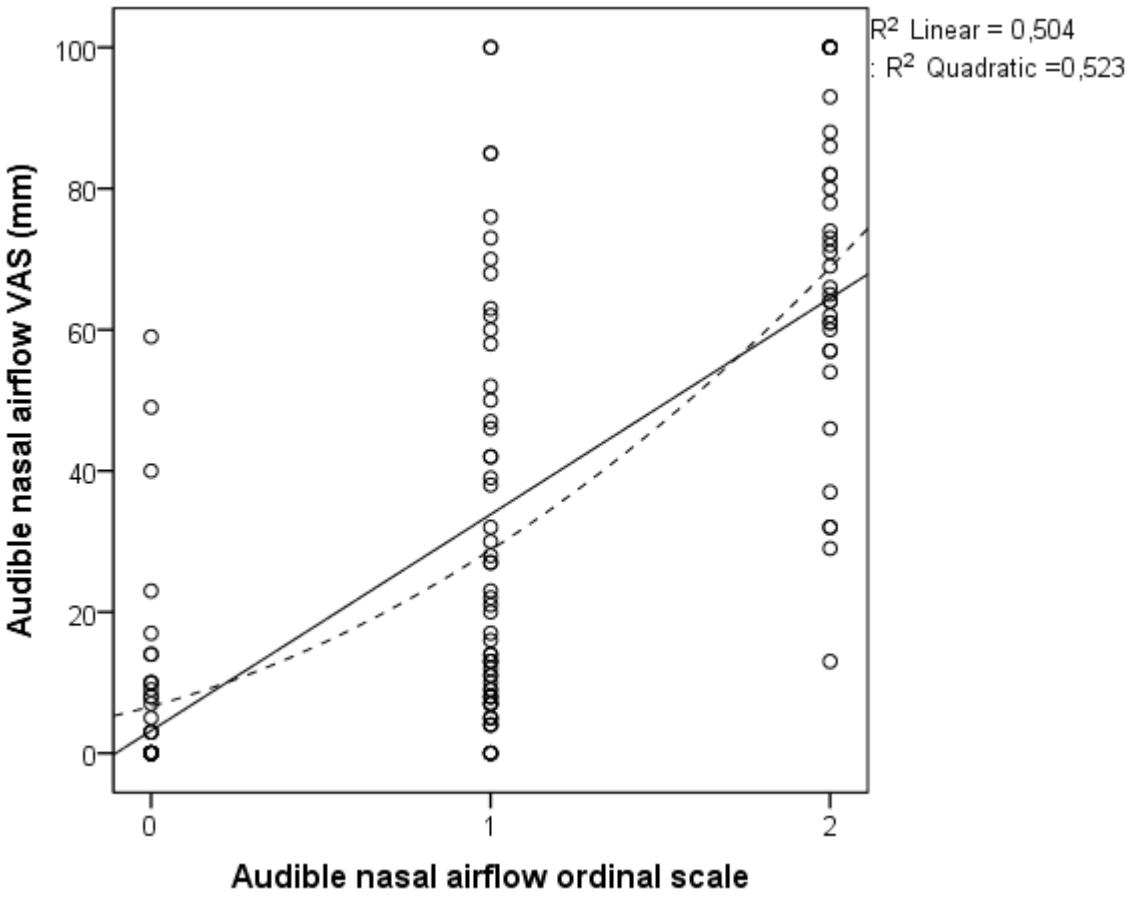

Figure 2. Individual audible nasal airflow (ANA) ratings based on visual analogue scale (VAS) plotted against individual ANA ratings based on an ordinal scale. Linear (y=3.194 + 30.639x) and curvilinear (y= 6.636 + 13.212x + 8.935x^2) relationships are provided.

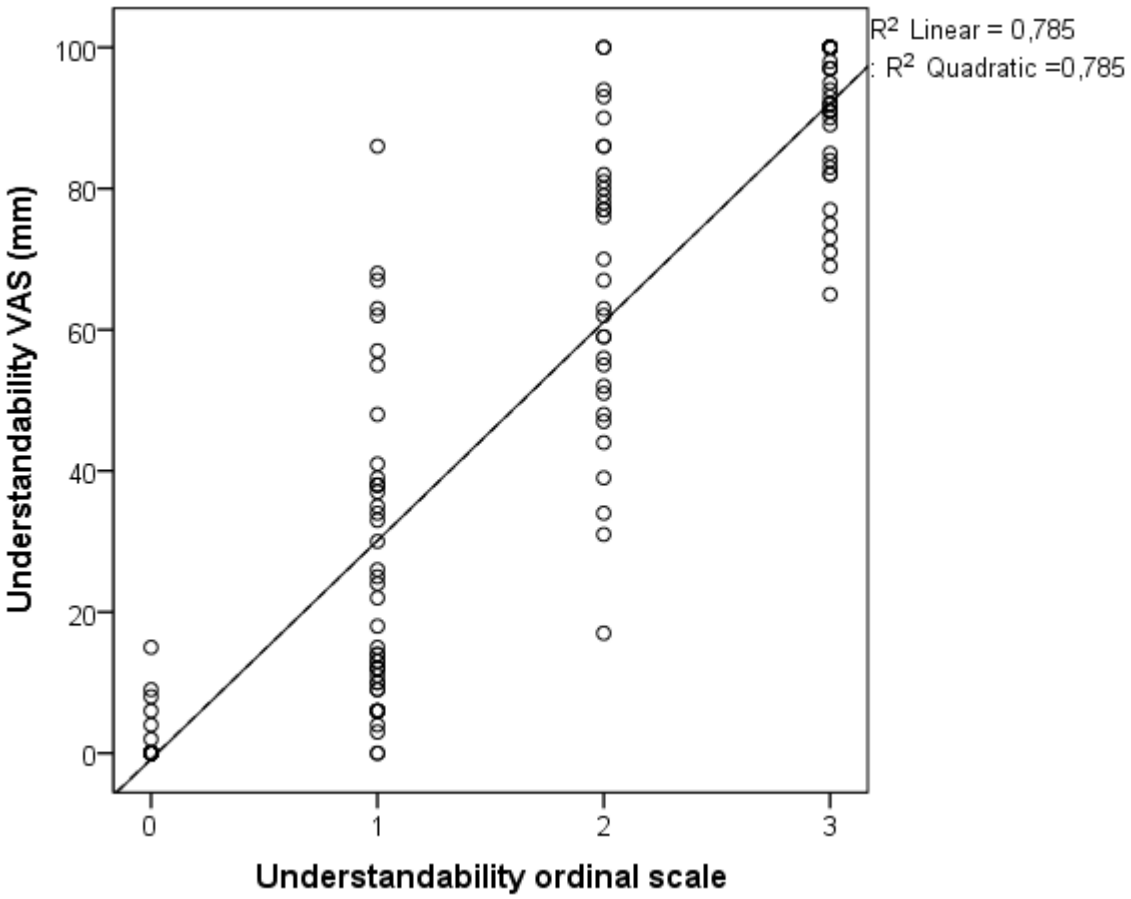

Figure 3. Individual ratings of understandability based on visual analogue scale (VAS) plotted against individual ratings of understandability based on an ordinal scale. Linear ($y=-0.849 + 30.988x$) and curvilinear ($y= -0.747 + 30.769x + 0.067x^2$) relationships are provided.