Acta Odontologica Scandinavica, 2014; 72: 120–129

informa
healthcare

ORIGINAL ARTICLE

Inter-rater and intra-rater agreement on the Nordic Orofacial Test—Screening examination in children, adolescents and young adults with cerebral palsy

SIV ELISABET EDVINSSON[1,2,3] & LARS-OLOV LUNDQVIST[2,3]

[1]*Child and Youth Habilitation Centre,* [2]*Centre for Rehabilitation Research, Örebro County Council, Örebro, Sweden, and* [3]*School of Health and Medical Sciences, Örebro University, Örebro, Sweden*

Abstract

Objective. To evaluate inter-rater and intra-rater agreement on the Nordic Orofacial Test–Screening (NOT-S) examination applied to children, adolescents and young adults with cerebral palsy (CP). ***Materials and methods.*** Using the NOT-S examination, two speech and language pathologists independently assessed video recordings of 48 subjects with CP aged 5–22 years and representing all CP sub-diagnoses and levels of gross motor function and manual ability. Thirty-one subjects were reassessed. Fifteen out of 17 items in the NOT-S examination domains (1) Face at rest, (2) Nose breathing, (3) Facial expression, (4) Masticatory muscle and jaw function, (5) Oral motor function and (6) Speech were rated using a 'yes' (dysfunction observed)/'no' format, generating an overall score of 0–6. ***Results.*** *Inter-rater agreement:* Twelve out of 15 items and five out of six domains showed acceptable unweighted kappa values ($\kappa = 0.46$–1.00). The lowest kappa value was found for domain 4 ($\kappa = -0.04$), although it had high inter-rater agreement (92%). The linear weighted kappa value for the overall NOT-S examination score was 0.65 (95% CI = 0.49–0.82). *Intra-rater agreement:* All items and domains showed acceptable unweighted kappa values (items 0.58–1.00 and 0.59–1.00, domains 0.81–1.00 and 0.62–0.89) for both raters. The linear weighted kappa value for the overall NOT-S examination score was 0.81 (95% CI = 0.63–0.99) for rater A and 0.54 (95% CI = 0.25–0.82) for rater B. ***Conclusions.*** The NOT-S examination has acceptable inter-rater and intra-rater agreement when used in young individuals with CP.

Key Words: *facial paralysis, speech disorders, observer variation, reproducibility of results*

Introduction

The Nordic Orofacial Test–Screening (NOT-S) is a comprehensive screening instrument for orofacial dysfunction, aimed to identify areas of orofacial dysfunction that require further evaluation [1]. The NOT-S was developed in a Nordic collaboration project, based on the need for a standard instrument for the many different health professions included in the work of providing healthcare for individuals with orofacial dysfunction [1]. After testing of the Swedish version of the NOT-S, it is now freely available in several other languages [2]. It is intended for health professionals, to use when difficulties in breathing, eating, speech or other orofacial functions are anticipated. Specific narrow band assessments exist [3–5], but, to our knowledge, the NOT-S is the only screening instrument applicable to a broad spectrum of orofacial dysfunctions and to individuals from the age of 3 years through the life span, irrespective of diagnosis.

Studies with the NOT-S have provided information on the prevalence of orofacial dysfunction in children with adenotonsillar hypertrophy [6], children and adults with ectodermal dysplasias [7], Prader–Willi syndrome [8] and Treacher Collins syndrome [9], as well as adults with Parkinson's disease [10]. To the best of our knowledge, no diagnosis-specific studies with the NOT-S have been performed concerning individuals with cerebral palsy (CP).

Cerebral palsy comprises a group of permanent, non-progressive conditions caused by central nervous lesions, damage or dysfunction developed during pregnancy or early childhood, resulting in physical

Correspondence: Siv Edvinsson, Centre for Rehabilitation Research, Box 1613, SE–701 16 Örebro, Sweden. Tel: +46 19 602 58 91. Fax: +46 19 25 55 59. E-mail: siv.edvinsson@orebroll.se

(Received 23 November 2012; accepted 27 May 2013)

ISSN 0001-6357 print/ISSN 1502-3850 online
DOI: 10.3109/00016357.2013.810771

disabilities. In addition to the more general motor and posture problems, CP is often associated with disturbances of sensation, perception and cognition. Due to bilateral corticobulbar dysfunction, many children with CP have oral motor dysfunction, such as feeding or speech impairment of varying degree [11–14], which can have a significant impact on daily life. Screening of conditions associated with CP is recommended to identify areas of impairment in need of further, specific evaluation [12].

The NOT-S has the potential to be a useful screening test of orofacial dysfunction for individuals with CP, although its reliability for use in this diagnosis is unknown. So far, only one study has been published applying the NOT-S to individuals with CP [1] and, in that study, the six individuals with CP were not included in the inter-rater agreement estimation (personal communication).

The NOT-S consists of two parts, a structured interview on everyday performance of orofacial function and a clinical examination. The latter is based on health professionals' observations of subjects' responses to tasks performed upon request. Because the two parts of the NOT-S differ considerably (questionnaire-based interview vs visual and auditive assessment) they, consequently, probably have different agreement. It can be argued that the NOT-S examination could be a larger source of rater disagreement than the NOT-S interview, because the NOT-S interview has a verbal 'yes'/'no' response format that limits the extent of raters' disagreement on interpretation of the interviewee's response. The impact of the two NOT-S parts separately on agreement measures is not known. Hence, more information is needed about the quality of the NOT-S examination, particularly the degree to which assessments are identical between raters (interrater agreement) or are repeated by raters (intrarater agreement) [15].

Therefore, the aim of the present study was to evaluate the inter-rater and intra-rater agreement of the NOT-S examination applied to children, adolescents, and young adults with CP.

Materials and methods

Subjects

The present study was based on data from a larger epidemiological, total population study in Örebro County examining functional ability in children, adolescents and young adults with CP. The young individuals were identified through the medical records of the Child and Youth Habilitation Centre in Örebro County Council, Örebro, Sweden. Individuals who were 4–21 years of age on November 1st, 2008, with a CP diagnosis according to the International Statistical Classification of Diseases and Related Health Problems, Tenth Revision (ICD-10), classification G 80.1–80.9 [16] and living in Örebro County, were examined for eligibility. No exclusion criteria were applied. A medical secretary identified eligible individuals and made a printout, from the medical records, of patients' birth date, gender and CP sub-diagnosis, made by paediatric neurologists according to the ICD-10. The first author (S.E.) gathered information of requirement for an interpreter, if needed. The individuals were informed of the study by mail and invited to participate. The data collection was conducted between May 2009 and May 2010. Among the 132 individuals eligible for this study, 129 subjects agreed to participate in the NOT-S interview and the collection of data on demographics and functional ability, 52 of whom also agreed to participate in the NOT-S examination procedure. Four subjects were not included in the statistical analyses in the present study as they were used as subjects during the training of the raters. Moreover, based on ethical grounds, video recordings of eight subjects were discontinued after domain 3. Consequently, the data on domain 1–3 is based on 48 subjects (27 male, 21 female) in the age range of 5.17–22.42 years (mean = 14.38 years, standard deviation (SD) = 4.81) and the data on domain 4–6 is based on 40 subjects of 5.70–22.42 years old (mean = 14.38 years, SD = 4.80). All levels of gross motor function and manual ability and all CP sub-diagnoses were represented (Table I).

Ethical approval was given by the Regional Ethical Review Board in Uppsala, Sweden (Dnr 2008/228). Informed child and adolescent assent and written, informed parental consent to the research procedures and publication of results were obtained. Young adults aged 18 years and older gave their own consent where possible.

Raters

Two speech and language pathologists independently performed the NOT-S examination assessments. One had been introduced to the NOT-S during basic education. The other had no previous experience of the NOT-S. They had 11 and 18 months, respectively, of clinical experience of working with young individuals with disabilities, including CP, and also of structured instrument testing. The raters had no information on the subjects' sub-diagnosis, functional abilities or demographic data except for two and four of the subjects, respectively (where the rater was the subjects' therapist). The raters were aware that their assessments would be compared and were, therefore, asked to avoid exchange of information regarding the subjects during the rating phase. After the initial assessment of all subjects, the raters were asked to perform a second assessment so as to enable evaluation of intra-rater agreement. In other words, they

Table I. Characteristics of the subjects (n = 48).

Characteristic	n (%)
Age groups	
Children, 5–12 years	17 (35)
Adolescents, 13–18 years	23 (48)
Young adults, 19–22 years	8 (17)
Gender	
Male	27 (56)
Female	21 (44)
Diagnosis, ICD-10	
G80.0 Spastic quadriplegic CP	1 (2)
G80.1 Spastic diplegic CP	15 (31)
G80.2 Spastic hemiplegic CP	10 (21)
G80.3 Dyskinetic CP	10 (21)
G80.4 Ataxic CP	4 (8)
G80.8 Other CP, mixed syndromes	4 (8)
G80.9 CP, unspecified	4 (8)
Gross motor function level, GMFCS-E&R	
I. Walks without limitations	13 (27)
II. Walks with limitations	11 (23)
III. Walks using a handheld mobility device	7 (15)
IV. Self-mobility with limitations; may use powered mobility	5 (10)
V. Transported in a manual wheelchair	12 (25)
Manual ability level, MACS	
I. Handles objects easily and successfully	9 (19)
II. Handles most objects but with somewhat reduced quality and/or speed of achievement	15 (31)
III. Handles objects with difficulty; needs help to prepare and/or modify activities	12 (25)
IV. Handles a limited selection of easily managed objects in adapted situations	1 (2)
V. Does not handle objects and has severely limited ability to perform even simple actions	11 (23)

CP, cerebral palsy; GMFCS-E&R, Gross Motor Function Classification System Expanded and Revised; ICD-10, International Statistical Classification of Diseases and Related Health Problems, 10th revision; MACS, Manual Ability Classification System.

were initially not aware that they would be asked to perform the assessments twice. This was done to minimize efforts to try to memorize the first assessments.

Measures

Orofacial function. As previously mentioned, the NOT-S consists of two parts: a structured interview and a clinical examination [2]. Each part covers six domains. The NOT-S total score (interview and examination pooled), applied to a clinic referred sample of patients within nine ICD-10 classification chapters and a control sample, has adequate sensitivity (0.96), specificity (0.63) and inter-rater agreement, with an unweighted kappa value range of 0.42–0.44 and a percentage agreement of 85% [1]. The six domains of the NOT-S examination consist of: (1) Face at rest (four items); (2) Nose breathing (one item); (3) Facial expression (three items); (4) Masticatory muscle and jaw function (two items); (5) Oral motor function (four items); and (6) Speech (three items). For a description of the items included in the present study, see Table II. Each item is assessed 'yes', meaning that the criterion of dysfunction is fulfilled, or 'no', meaning there is no dysfunction. If one or more items within a domain are assessed with 'yes', this yields 1 point. The score of the NOT-S examination can vary between 0–6 points.

In the present study, some modifications of the NOT-S examination were made. First, the NOT-S examination was recorded on video, to be scored later. The use of recordings ensured that both raters assessed the same responses seen from the same angle. Secondly, based on clinical experience, a criterion for a maximum time to respond was added to the items of domains 2–6. The maximum response time was set at 90 s after verbal instructions illustrated by photos had been given. The instructions were repeated twice and further clarified within the response time. Thirdly, item 4A (Masticatory muscle and jaw function—Bite hard on your back teeth) was not assessed because it is based on manual palpation of the masseter muscles, which was not possible to perform given the present study procedure. Fourthly, item 5D (Oral motor function—Open your mouth wide and say 'ah-ah-ah') for elevation of the uvula and the soft palate was not possible to assess due to insufficient light in the mouth and/or the fixed placement of the camcorder. Hence, 15 out of 17 items were used in the analyses.

To reduce the strain on subjects unable to perform specific oral motor behaviors on demand, criteria for discontinuing the NOT-S examination were established. The NOT-S examination was discontinued if the subject did not make any attempt to respond to the instructions in domain 3 (Facial expression) and if the parent confirmed that their child was not able to follow instructions for orofacial activities. We made this decision on ethical grounds, although we knew that this would reduce the number of recorded items.

In accordance with the NOT-S screening form, item 2A, and consequently domain 2 (Nose breathing), was not assessed if the subject's nose was blocked by a cold. In addition, there were also single items occasionally not assessed because some subject could not sit still enough to make their face clearly visible or did not co-operate or the recording was too short. The number of eligible assessments in each estimation of agreement is presented in Tables II and III.

Table II. Inter-rater agreement on the Nordic Orofacial Test–Screening (NOT-S) examination. Unweighted kappa values with 95% confidence intervals (CIs), prevalence index, bias index, maximum attainable unweighted kappa value (κ_{max}) and percentage agreement results are given.

NOT-S examination domain (numeral) and item (letter)	Subjects (*n*)	Kappa (95% CI)	Prevalence index[a]	Bias index[b]	κ_{max}[c]	Agreement (%)
1 – Face at rest	37	0.66 (0.42; 0.90)	0.24	0.11	0.77	84
1A Asymmetry	36	0.09 (−0.20; 0.38)	0.44	0.22	0.48	61
1B Deviant lip position	37	0.95 (0.84; 1.05)	0.11	0.03	0.95	97
1C Deviant tongue position	37	0.72 (0.42; 1.02)	0.65	0.03	0.91	92
1D Involuntary movements	37	0.47 (0.09; 0.86)	0.70	0.08	0.68	87
2 – Nose breathing: 2A Close your mouth and take five deep breaths through your nose (smell)	27	0.61 (0.27; 0.96)	0.48	0.00	1.00	85
3 – Facial expression	39	0.46 (0.15; 0.78)	0.49	0.05	0.93	80
3A Close your eyes tightly	36	0.55 (0.28; 0.82)	0.11	0.06	0.89	78
3B Show your teeth	38	0.73 (0.52; 0.95)	0.13	0.03	0.95	88
3C Try to whistle (blow)	36	0.89 (0.74; 1.04)	0.06	0.00	1.00	94
4 – Masticatory muscle and jaw function: 4B Open your mouth as wide as you can	36	−0.04 (−0.09; 0.01)	0.92	0.03	0.65	92
5 – Oral motor function	35	0.51 (0.24; 0.78)	0.31	0.17	0.63	77
5A Stick out your tongue as far as you can	37	1.00 (1.00; 1.00)	0.68	0.00	1.00	100
5B Lick your lips	34	0.35 (0.07; 0.64)	0.21	0.21	0.59	68
5C 'Blow up' your cheeks and hold for at least 3 s	34	0.70 (0.45; 0.94)	0.21	0.09	0.82	85
6 – Speech	34	0.61 (0.34; 0.89)	0.29	0.00	1.00	82
6A Does not speak	26	1.00 (1.00; 1.00)	0.62	0.00	1.00	100
6B Count loud to ten	27	0.69 (0.42; 0.96)	0.26	0.15	0.69	85
6C Say 'pataka, pataka, pataka'	26	0.70 (0.44; 0.96)	0.08	0.15	0.70	85

[a]Prevalence index is the absolute value of the difference between the frequencies of agreed 'yes' assessments and agreed 'no' assessments, divided by the number of paired ratings.
[b]Bias index is the absolute value of the difference between the frequency of Rater A's 'no' assessments when Rater B assess 'yes' and the frequency of Rater B's 'no' assessments when Rater A assess 'yes', divided by the number of paired ratings.
[c]κ_{max} is the maximum attainable unweighted kappa value calculated with the marginal totals taken as fixed and the cell frequencies adjusted to represent the greatest possible agreement.

Classification of gross motor function and manual ability. Gross motor function was classified according to the expanded and revised Gross Motor Function Classification System (GMFCS-E&R) [17,18] and manual ability according to the Manual Ability Classification System (MACS) [19,20]. The Swedish version of both of these classification systems was used. For both instruments, the classification consists of five levels, from level I = most able, to level V = most limited. For descriptions of the levels, see Table I. The parent of the subject or, in some cases, the young adult him/herself determined what level best described the usual performance. As there currently are no descriptive criteria for young adults, the age band up to 18 was used to classify also those older than 18 years. The classifications were used to compare the gross motor function and manual ability of the subjects in the present study with the total population and to evaluate if the raters' assessments were influenced by gross motor function level or manual ability level.

Procedure

The NOT-S examination was conducted in a habitual environment chosen by the parents and their child, usually in their home. It was carried out, according to the screening form [2], in Swedish and, for six individuals, with an interpreter present, because, at the time of the study, the NOT-S was not available in any language they could understand. The first author (S.E.) performed all NOT-S examinations as the subjects were video-recorded with a memory camcorder placed at face level on a tripod in front of the subject. The video recordings were then transferred to two sets of DVD discs in random order. The first set was used for estimation of inter-rater agreement and both for intra-rater agreement estimation. The unit for randomization was the complete recording of each subject. In the set used for intra-rater agreement estimation, a stratified randomization based on the gross motor function level of the subjects was made.

Table III. Intra-rater agreement on the Nordic Orofacial Test–Screening (NOT-S) examination. Unweighted kappa values with 95% confidence intervals (CIs), prevalence index, maximum attainable unweighted kappa value (κ_{max}) and percentage agreement results are given.

	Rater A					Rater B				
NOT-S examination domain (numeral) and item (letter)	Subjects (*n*)	Kappa (95% CI)	Prevalence index[a]	κ_{max}[b]	Agreement (%)	Subjects (*n*)	Kappa (95% CI)	Prevalence index	κ_{max}	Agreement (%)
1 – Face at rest	26	1.00 (1.00; 1.00)	0.15	1.00	100	23	0.89 (0.69; 1.10)	0.44	0.89	96
1A Asymmetry	25	0.63 (0.17; 1.09)	0.76	0.63	92	23	0.62 (0.29; 0.95)	0.30	0.81	83
1B Deviant lip position	26	1.00 (1.00; 1.00)	0.07	1.00	100	23	0.83 (0.60; 1.05)	0.04	0.91	91
1C Deviant tongue position	26	1.00 (1.00; 1.00)	0.77	1.00	100	23	0.78 (0.36; 1.19)	0.78	0.78	96
1D Involuntary movements	26	0.65 (0.26; 1.03)	0.89	0.65	96	23	0.59 (0.17; 1.00)	0.61	0.86	87
2 – Nose breathing: 2A Close your mouth and take five deep breaths through your nose	20	1.00 (1.00; 1.00)	0.30	1.00	100	19	0.86 (0.58; 1.13)	0.53	0.86	100
3 – Facial expression	28	0.81 (0.56; 1.06)	0.50	0.81	93	25	0.71 (0.33; 1.08)	0.68	0.71	92
3A Close your eyes tightly	26	0.84 (0.62; 1.05)	0.23	1.00	92	23	0.63 (0.32; 0.94)	0.30	0.63	83
3B Show your teeth	26	1.00 (1.00; 1.00)	0.15	1.00	100	24	0.74 (0.48; 1.00)	0.21	0.74	88
3C Try to whistle	28	0.86 (0.67; 1.05)	0.07	0.86	93	24	0.92 (0.76; 1.08)	0.04	0.92	96
4 – Masticatory muscle and jaw function: 4B Open your mouth as wide as you can	24	1.00 (1.00; 1.00)	0.92	1.00	100	23	0.62 (0.16; 1.09)	0.74	0.62	91
5 – Oral motor function	23	0.83 (0.52; 1.15)	0.70	0.83	96	23	0.73 (0.45; 1.00)	0.26	0.73	87
5A Stick out your tongue as far as you can	24	1.00 (1.00; 1.00)	0.75	1.00	100	24	0.83 (0.52; 1.15)	0.71	0.83	96
5B Lick your lips	22	0.58 (0.16; 1.00	0.59	0.86	86	22	0.73 (0.46; 1.00)	0.05	0.73	86
5C Blow up your cheeks and hold for at least 3 s	22	0.81 (0.57; 1.06)	0.18	0.81	91	22	0.81 (0.57; 1.06)	0.18	0.81	91
6 – Speech	22	0.81 (0.56; 1.06)	0.18	1.00	91	21	0.81 (0.54; 1.07)	0.14	1.00	91
6A Does not speak	17	1.00 (1.00; 1.00)	0.53	1.00	100	16	1.00 (1.00; 1.00)	0.64	1.00	100
6B Count loud to ten	17	0.85 (0.57; 1.13)	0.47	0.85	94	18	1.00 (1.00; 1.00)	0.00	1.00	100
6C Say ‘pataka, pataka, pataka’	17	0.76 (0.46; 1.07)	0.06	1.00	88	16	0.71 (0.34; 1.08)	0.38	1.00	88

[a]Prevalence index is the absolute value of the difference between the frequencies of agreed ‘yes’ assessments and agreed ‘no’ assessments, divided by the number of paired ratings.
[b]κ_{max} is the maximum attainable unweighted kappa value calculated with the marginal totals taken as fixed and the cell frequencies adjusted to represent the greatest possible agreement.

Calibration. In accordance with previous studies [1,7,8], a training and calibration phase was undertaken before the assessments were made. The raters were first introduced to the NOT-S examination. Then, joint as well as individual assessments of video-recorded NOT-S examinations were followed by discussions to reach consensus on how to interpret the criteria of item fulfillment. Recordings of patients not included in this study, four with CP and five with other neurological disorders, were used for this purpose.

Assessment. After the calibration phase, the two raters independently viewed the recorded NOT-S examinations of the 48 subjects. One subject was assessed at a time. Thereafter, to evaluate intra-rater agreement, the two raters independently reassessed a random sample consisting of 31 of the 48 subjects. It was not possible to reassess all subjects because of extraneous restrictions on the raters' availability. The reassessments, performed 1 week after the initial assessments were done without access to previous results. However, both rater A and rater B missed assessing one subject in one item (different subjects and items). Finally, after completing all reassessments, the raters were informed of the inter-rater agreement results and interviewed by the first author (S.E.). The focus of these interviews was to discuss items with low values of agreement in order to explore possible reasons for discrepancy.

Statistical analysis

The present study is reported in accordance with the Guidelines for Reporting Reliability and Agreement Studies (GRRAS) [15]. Descriptive and inferential statistics were calculated using PASW/PC version for Windows 17.0 (IBM SPSS Inc., Chicago, IL). Kappa statistics and McNemar's test were calculated using VassarStats [21].

Kappa values were interpreted using Altman's [22] adaptation of Landis and Koch's [23] guidelines. The guidelines were used for unweighted as well as weighted kappa values, as recommended by Fleiss et al. [24]. A kappa value of 0 reflects an agreement expected by chance; kappa values of 0.01–0.20 are interpreted as 'poor', 0.21–0.40 as 'fair', 0.41–0.60 as 'moderate', 0.61–0.80 as 'good' and 0.81–1.00 as 'very good' [22]. For screening purposes, a kappa value ≥ 0.41 was considered to be sufficiently reliable.

The inter-rater and intra-rater agreement was estimated item-by-item and domain-by-domain, giving percentage agreement and unweighted kappa (κ) with 95% confidence intervals (CIs). Unweighted kappa was calculated on 4-fold matrices. Because kappa values are sensitive to prevalence and do not indicate if low values of agreement are due to random or systematic differences, the data was further examined. Firstly, the prevalence index [25] was calculated to determine the degree to which raters' proportion of agreements of 'yes' (dysfunction) differed from the proportion of agreements of 'no' (no dysfunction). Prevalence index values range from 0–1, where 0 indicates that the proportions have an even distribution [25]. Secondly, bias index [25] was calculated to determine the extent to which the raters disagreed on the proportion of 'yes'. Bias index values range from 0–1, where 0 indicates no difference in proportion of positive ratings between the raters [25]. Thirdly, because the distribution of the marginal totals in a 4-fold contingency table does not always allow kappa to reach a maximum of 1, the maximum attainable unweighted kappa value (κ_{max}) was calculated, with the marginal totals taken as fixed and the cell frequencies adjusted, to represent the greatest possible agreement [25].

McNemar's test for correlated proportions in the marginal results of a 4-fold table was used to check for significant differences between the raters. Chi-square (χ^2) analysis with Fisher's exact test was used to evaluate if the raters' assessments were influenced by subjects' age, gender, level of gross motor function or manual ability. A statistical significance level of $p < 0.05$ was used for all analyses.

The overall score for the NOT-S examination is ordinal and varies between 0–6. A difference of one unit was considered less serious than a difference of 2 or more units. Therefore, weighted kappa was considered. Because kappa with linear weighting is less sensitive to the number of categories compared to kappa with quadratic weighting [26], kappa with linear weighting with 95% CI was used to estimate inter-rater and intra-rater agreement of the overall NOT-S examination score. The maximum attainable linear weighted kappa value and the percentage agreement were calculated. To compare the results of the present study to results presented in the study of Bakke et al. [1], we also calculated unweighted kappa for the overall NOT-S examination score.

Results

There were no significant differences concerning age, gender, CP sub-diagnoses or level of gross motor function between the subjects participating in the NOT-S examination ($n = 48$) and those not participating ($n = 81$). There was, however, a significant difference in MACS level, with the NOT-S examination sub-sample having a larger proportion of individuals with less manual ability ($\chi^2 = 14.05$, df = 4; $p = 0.006$), indicating a skewness towards more manual disability in the NOT-S examination sub-sample. However, eight subjects discontinued the recordings after domain 3. When the analyses were repeated with

the remaining sub-sample (n = 40) compared to the sub-sample of 81 subjects not participating in the examination, no significant differences were revealed. All eight subjects who participated in domains 1–3 only had severe limitations in manual ability and gross motor function (levels IV–V).

Inter-rater agreement

Item level. The unweighted kappa value of the items ranged from −0.04 to 1.00, the prevalence indices varied between 0.06–0.92 and the bias indices varied between 0.00–0.22. The percentage agreement ranged from 61–100% (Table II). All items but three, 1A (Face at rest—Asymmetry), 4B (Masticatory muscle and jaw function—Open your mouth as wide as you can) and 5B (Oral motor function—Lick your lips), had at least moderate unweighted kappa values ($\kappa \geq 0.41$). Items 1A and 4B had kappa values close to zero. The low kappa value for 4B was accompanied with a high prevalence index (0.92), indicating a large imbalance between the number of subjects where both raters agreed on the subject's ability to open their mouth (n = 33) and the number of subjects where inter-rater agreement was reached on their inability to do so (n = 0). This was not the case for items 1A and 5B. In fact, these items had the highest bias indices among all items (0.22 and 0.21, respectively) and the lowest percentage agreement (61% and 68%, respectively), indicating low inter-rater agreement. However, no rater had a significantly larger proportion of 'yes' or 'no' assessments (all p-values > 0.05), indicating no systematic difference between the raters.

There were no significant differences between the raters' assessments in relation to the gender or level of gross motor function of subjects. There was, however, a significant difference between the raters' assessment of item 5C (Oral motor function—Blow up your cheeks and hold for at least 3 s). The raters agreed significantly more on subjects in the older, 15–22-years, age group (agreement = 100%) than on subjects in the younger, 5–14-years, age group (agreement = 62%), as indicated by a chi-square result of 9.47 (df = 1; p = 0.046). Differences in rater agreement were also found for item 6C (Speech—Say 'pataka, pataka, pataka') in relation to manual ability. That is, the raters' assessments agreed significantly more on (more manually able) subjects at MACS levels I–II (agreement = 95%) than on subjects at MACS levels III–V (agreement = 57%), as indicated by a chi-square result of 5.55 (df = 1; p = 0.047).

Domain level. All domains except domain 4 (Masticatory muscle and jaw function) had a moderate unweighted kappa value ranging from 0.46–0.66 (Table II). The prevalence indices ranged from 0.24–0.92 and the bias indices ranged from 0.00–0.17. The maximum attainable unweighted kappa value ranged from 0.63–1.00 and the percentage agreement ranged from 77– 92%. There were no significant differences between the raters' proportions of 'yes' and 'no' assessments in any domain (all p-values > 0.07).

Inter-rater agreement in items and domains without an interpreter. In the inter-rater agreement estimations on recordings made without an interpreter (n = 42), the unweighted kappa values in items and domains differed from the estimations on all 48 subjects by −0.03 to +0.09. However, this only affected the interpretation of domain 6—Speech, having an unweighted kappa value of 0.61 (good) in the sub-sample with interpreters present compared to 0.58 (moderate) in the sub-sample with 42 subjects and no interpreter present.

NOT-S examination score. The overall NOT-S examination score was determined for the 21 subjects with assessable recordings in all six domains. The linear weighted kappa value of the overall NOT-S examination score was 0.65 (95% CI = 0.49–0.82). The maximum attainable linear weighted kappa value was 0.78 and the percentage agreement was 52%. For comparative reasons, the unweighted kappa value was estimated and found to be 0.40 (95% CI = 0.14–0.66).

Intra-rater agreement

Item level. As shown in Table III, the unweighted kappa value of the items ranged from 0.58–1.00 and from 0.59–1.00 for the two raters, respectively. This was interpreted as moderate-to-very good agreement ($\kappa \geq 0.41$). The prevalence indices of the items ranged from 0.06–0.92 and from 0.04–0.78, respectively. The maximum attainable unweighted kappa values ranged from 0.63–1.00 and from 0.62–1.00, respectively. The percentage agreement ranged from 86–100% and from 83–100%, respectively.

Domain level. At the domain level, all unweighted kappa values indicated a level of good agreement. The unweighted kappa ranged from 0.81–1.00 and from 0.62–0.89, respectively, for the two raters (Table III). The prevalence indices ranged from 0.15–0.92 for rater A and from 0.14–0.74 for rater B. The maximum attainable unweighted kappa values ranged from 0.81–1.00 and from 0.62–1.00, respectively. The percentage agreement ranged from 91–100% and from 87–100%, respectively.

NOT-S examination score. The overall NOT-S examination score was determined for 12 subjects assessed a second time in all six domains by both raters. The linear weighted kappa value for the NOT-S examination score was 0.81 (95% CI = 0.63–0.99) for rater A and 0.54 (95% CI = 0.25–0.82) for rater B. Both raters reached the maximum attainable linear weighted kappa value given the observed marginal frequencies. The percentage agreement was 75% and 42%, respectively. The unweighted kappa values were 0.67 (95% CI = 0.36–0.98) and 0.26 (95% CI = 0.00–0.60), respectively.

Discussion

The results demonstrate acceptable inter-rater agreement in 12 out of 15 items, in five out of six domains and for the overall NOT-S examination score. The intra-rater agreement was adequate for both raters. At the item level, both raters achieved moderate-to-very good intra-rater agreement. At the domain level, one of the raters achieved very good agreement and the other had good-to-very good agreement. The intra-rater agreement on the overall NOT-S examination score was very good and moderate, respectively.

This is, to our knowledge, the first study to evaluate inter-rater and intra-rater agreement on the NOT-S examination applied to children, adolescents and young adults with CP and also to estimate the agreement at the item and domain levels. Moreover, it is the first evaluation of agreement using raters similar to the intended users of the NOT-S, thus resembling a clinical situation.

The inter-rater agreement on the NOT-S examination in the present study was acceptable. However, there was a tendency for a significant difference between the raters in two of the items, 1A and 5B. When we interviewed the raters after the rating phase, they expressed different opinions on how much asymmetry was needed to rate the face as asymmetric in item 1A (Face at rest—Asymmetry). In item 5B (Oral motor function—Lick your lips), the two raters differed in opinion on the degree of unevenness of the tongue movement when wetting the lips that was required for the item to be rated as dysfunctional. According to Bakke et al. [1], raters should be trained and calibrated before performing screening. This was done in the present study, but the above results illustrate difficulty in dichotomizing continuous variables. It is possible that clearer criteria and structured training would increase agreement between raters on these items.

For comparison reasons, unweighted kappa values on the overall NOT-S examination score were estimated. In the original study on the NOT-S, by Bakke et al. [1], the inter-rater agreement was estimated on the basis of the pooled NOT-S interview and examination scores (12 domains). In the present study, the score was based on the six domains of the examination part of the NOT-S. The estimated unweighted kappa value in the present study was 0.40, which is in line with the value reported by Bakke et al. [1] for the total score for the pooled NOT-S, suggesting that the NOT-S examination is useful for children, adolescents and young adults with CP. Recently, Åsten et al. [9] reported inter-rater agreement on the pooled NOT-S, applied to patients with Treacher Collins syndrome. Their linear weighted kappa value was 0.75, which is somewhat higher than in the present study (0.65).

One of the strengths of the present study is that it was conducted on a population-based sample including all eligible subjects, regardless of their level of cognition or spoken language, enhancing the generalizability of this study. In addition, subjects with any CP sub-diagnosis and any level of gross motor function and manual ability were included, regardless of orofacial ability. This resulted in the overall NOT-S examination score ranging across almost the full range of possible scores (i.e. 0–5, compared to the maximum range of 0–6).

Some limitations of the present study should be noted. Firstly, the number of subjects that could be assessed in all of the six domains was low (44% of the eligible subjects). This was a consequence of including subjects with severe CP and adopting criteria for discontinuation of the examination procedure, based on ethical considerations. This has affected mainly the inter-rater and intra-rater agreement estimation of the overall NOT-S examination score. However, this is the largest NOT-S study on subjects with CP that has been performed so far.

A second limitation is related to the use of video recordings. Although video recording was necessary for estimation of intra-rater agreement, the use of video-recorded examinations instead of *in vivo* examinations restricted adjustments of visual angle and response time adapted to subjects, which may have led to an under-estimation of the rater agreement compared to examinations done *in vivo*, which would be the normal clinical procedure. Increasing the number of video cameras using different angles may augment information. Nevertheless, two of the items, activity of the masseter muscles and movements of the uvula and soft palate, are not readily accessible with video recordings alone. To some extent, the movements of the uvula and soft palate can be assessed by means of a hand-held, close-up camera, with additional light directed into the mouth.

Thirdly, the standardized response time limitation adopted in this study may have given false positive assessments because of subjects' prolonged time to respond as a consequence of their brain damage, although one could argue that, after a certain length

of time, a response made would be considered dysfunctional.

In addition to study-related limitations, there were some limitations linked to the NOT-S examination screening form [2]. Interpretations of the criteria for dysfunction were the main basis for rater disagreements. This warrants clearer definitions. For example, a reference and training manual and CD-ROM with more explicit and illustrated criteria, including, e.g. illustration of face asymmetry, would probably increase NOT-S rater agreement.

The NOT-S examination assesses orofacial ability in tasks performed on demand. Normally the NOT-S includes an interview that adds information about orofacial functions in daily life. Multidimensional screening instruments, such as the NOT-S, are not aimed to give detailed information but are useful when they precede specific evaluations or classifications. Currently, classifications for children with CP describing orofacial function in daily activities, designed in the same five level formats as the instruments for gross motor function (GMFCS-E&R) and manual ability (MACS) and involving some areas of orofacial function assessed in the NOT-S, are under development [27–29].

Hitherto, all agreement studies on the NOT-S have been conducted with video-recorded examinations [1,6,9]. Because the NOT-S is designed for assessments made *in vivo*, studies need to evaluate the NOT-S using *in vivo* assessments. In addition, comparisons of assessments made *in vivo* vs video recordings are needed. Furthermore, studies evaluating the inter-rater and intra-rater agreement on the NOT-S have until the present been based on a small number of subjects. To establish the quality of the NOT-S in terms of agreement and reliability, further research is needed on larger samples of subjects with a definite diagnosis and preferably with more raters from different professions structurally trained in using the NOT-S.

In conclusion, the study shows that the NOT-S examination has acceptable inter-rater and intra-rater agreement when applied to children, adolescents and young adults with CP. Health professionals with limited experience and training in the NOT-S examination can use it, performing their assessment from video recordings. Although clearer criteria and structured training in the NOT-S instrument would improve its usability, the NOT-S examination in its present form can be reliably performed in young individuals with CP.

Acknowledgment

Financial support for the study has been gratefully received from the Folke Bernadotte Foundation, the Norrbacka-Eugenia Foundation, the Samariten Foundation and the Örebro County Council Research Committee.

Declaration of interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

References

[1] Bakke M, Bergendal B, McAllister A, Sjögreen L, Åsten P. Development and evaluation of a comprehensive screening for orofacial dysfunction. Swed Dent J 2007;31:75–84.

[2] Bakke M, Bergendal B, McAllister A, Sjögreen L, Åsten P. Nordic Orofacial Test – Screening, NOT-S. Available online at http://mun-h-center.se/en/Mun-H-Center/Mun-H-Center-E/NOT-S/.accessed March 5, 2013.

[3] Santos MT, Manzano FS, Ferreira MC, Masiero D. Development of a novel orofacial motor function assessment scale for children with cerebral palsy. J Dent Child 2005; 72:113–18.

[4] Ortega AOL, Ciamponi AL, Mendes FM. Assessment scale of the oral motor performance of children and adolescents with neurological damages. J Oral Rehabil 2009;36:653–9.

[5] Sonies BC, Cintas HL, Parks R, Miller J, Caggiano C, Gottshall SG, et al. Brief assessment of motor function: content validity and reliability of the oral motor scales. Am J Phys Med Rehabil 2009;88:464–72.

[6] Lundeborg I, McAllister A, Graf J, Ericsson E, Hultcrantz E. Oral motor dysfunction in children with adenotonsillar hypertrophy – effects of surgery. Log Phon Voc 2009;34: 111–16.

[7] Bergendal B, McAllister A, Stecksén-Blicks C. Orofacial dysfunction in ectodermal dysplasias measured using the Nordic Orofacial Test – Screening protocol. Acta Odontol Scand 2009;67:377–81.

[8] Saeves R, Åsten P, Storhaug K, Bågesund M. Orofacial dysfunction in individuals with Prader-Willi syndrome assessed with NOT-S. Acta Odontol Scand 2011;69:310–15.

[9] Åsten P, Skogedal N, Nordgarden H, Axelsson S, Akre H, Sjögreen L. Orofacial functions and oral health associated with Treacher Collins syndrome. Acta Odontol Scand 2013; 71:616–25.

[10] Bakke M, Larsen SL, Lautrup C, Karlsborg M. Orofacial function and oral health in patients with Parkinson's disease. Eur J Oral Sci 2011;119:27–32.

[11] Sullivan PB, Lambert B, Rose M, Ford-Adams M, Johnson A, Griffiths P. Prevalence and severity of feeding and nutritional problems in children with neurological impairment: Oxford feeding study. Dev Med Child Neurol 2000;42:674–80.

[12] Ashwal S, Russman BS, Blasco PA, Miller G, Sandler A, Shevell M, et al. Practice parameter: diagnostic assessment of the child with cerebral palsy: report of the Quality Standards Subcommittee of the American Academy of Neurology and the Practice Committee of the Child Neurology Society. Neurology 2004;62:851–63.

[13] Parkes J, Hill N, Platt MJ, Donelly C. Oromotor dysfunction and communication impairments in children with cerebral palsy: a register study. Dev Med Child Neurol 2010;52:1113–19.

[14] Dahlseng MO, Finbråten AK, Júlíusson PB, Skarnes J, Andersen G, Vik T. Feeding problems, growth and nutritional status in children with cerebral palsy. Acta Paediatr 2012;101: 92–8.

[15] Kottner J, Audigé L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, et al. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. J Clin Epidemiol 2011;64:96–106.

[16] World Health Organization. International Statistical Classification of Diseases and Related Health Problems: ICD-10. 10. rev, 2008 ed. Geneva: World Health Organization; 2009.
[17] Palisano RJ, Rosenbaum P, Bartlett D, Livingston MH. Content validity of the expanded and revised Gross Motor Function Classification System. Dev Med Child Neurol 2008;50: 744–50.
[18] Palisano R, Rosenbaum P, Bartlett D, Livingston M. GMFCS – E & R Gross Motor Function Classification System Expanded and Revised. Available online at http://www.canchild.ca/en/measures/gmfcs_expanded_revised.asp.accessed March 5, 2013.
[19] Eliasson AC, Krumlinde-Sundholm L, Rösblad B, Beckung E, Arner M, Öhrvall AM, et al. The Manual Ability Classification System (MACS) for children with cerebral palsy: scale development and evidence of validity and reliability. Dev Med Child Neurol 2006;48:549–54.
[20] Eliasson AC, Krumlinde-Sundholm L, Rösblad B, Beckung E, Arner M, Öhrvall AM, et al. The Manual Ability Classification System (MACS). 2005. Available online at http://www.macs.nu.accessed March 5, 2013.
[21] VassarStats. Website for Statistical Computation. Available online at http://vassarstats.net.accessed March 5, 2013.
[22] Altman DG. Practical statistics for medical research. London: Chapman &Hall/CRC; 1999.
[23] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.
[24] Fleiss JL, Levin B, Paik MC. Statistical methods for rates and proportions. 3rd ed. Hoboken, NJ: Wiley; 2003.
[25] Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys Ther 2005;85:257–68.
[26] Brenner H, Kliebsch U. Dependence of weighted kappa coefficients on the number of categories. Epidemiology 1996;7:199–202.
[27] Andrada M, Virella D, Gouveia R, Calado E, Folha T, Andrada M. Validation of assessment scales for communication and oro-motor function in children with cerebral palsy. Dev Med Child Neurol 2008;50:29.
[28] Hidecker MJC, Paneth N, Rosenbaum P, Kent RD, Lillie J, Eulenberg JB, et al. Developing and validating the Communication Function Classification System for individuals with cerebral palsy. Dev Med Child Neurol 2011;53:704–10.
[29] Sellers D, Pountney T, Pennington L, Morris C, Mandy A, Hankins M. Development of a functional classification system of eating and drinking ability for individuals with cerebral palsy. Dev Med Child Neurol 2012;54:8.