Central European Journal of Nursing and Midwifery

Cent Eur J Nurs Midw 2016;7(2):412–419
doi: 10.15452/CEJNM.2016.07.0009

ORIGINAL PAPER

INTER-RATER AGREEMENT OF THE BRIEF BEDSIDE DYSPHAGIA SCREENING TEST-REVISED IN PATIENTS WITH STROKE

Petra Mandysová[1,2], Helena Trundová[1], Edvard Ehler[2,3]

[1]Department of Nursing, Faculty of Health Studies, University of Pardubice, Czech Republic
[2]Department of Neurology, Pardubice Hospital, Hospitals of the Pardubice Region, Czech Republic
[3]Department of Clinical Subspecialties, Faculty of Health Studies, University of Pardubice, Czech Republic

Received October 17, 2015; Accepted November 18, 2015.

Abstract

Aim: The aim of the study was to investigate the inter-rater agreement (IRA) of the 8-item Brief Bedside Dysphagia Screening Test-Revised (BBDST-R). *Design:* An observational IRA study was conducted. *Methods:* Forty-six patients with stroke were independently assessed by two nurse raters using the BBDST-R. Rater agreement was described and analysed using descriptive statistics, Cohen's kappa (κ), the proportion of observed agreement (P_o), the prevalence index (p_{index}), and the bias index (b_{index}). *Results:* Kappa ranged from -0.046 (p = 0.641; 95% CI = -0.126–0.034) for item "ability to clench the teeth" to 0.784 (p < 0.001; 95% CI = 0.377–1.000) for "thick liquid: cough". For the overall BBDST-R result, κ was 0.241 (p = 0.114; 95% CI = 0.000–0.558). The P_o was ≥ 0.80 for five items; the p_{index} was ≥ 0.80 for three items. The b_{index} ranged from 0.03–0.31. *Conclusion:* The IRA, as expressed by κ, was low for the overall result and variable for the individual items. However, for some items, κ may have misrepresented the IRA due to the high p_{index} and b_{index}. Several strategies are recommended to improve the IRA of the instrument. This should enhance the instrument's capacity to produce consistent results across raters.

Keywords: Brief Bedside Dysphagia Screening Test-Revised, inter-rater agreement, dysphagia screening, nurse, stroke.

Introduction

Dysphagia is a common complication of stroke that may lead to serious consequences, such as malnutrition, dehydration, aspiration, and aspiration pneumonia (Martino et al., 2012). Because care provided in the first hours after stroke is critical in shaping patients' long-term recovery and prognosis, rapid initial assessment is essential (Middleton et al., 2015). Nurses play a pivotal role in this early assessment and the identification of problem areas that warrant further clinical investigation by other members of the multidisciplinary team (Middleton et al., 2015).

One of the important decision points involves determining whether patients with stroke appear to be suffering from swallowing difficulties; to this end, various bedside screening tests have been developed (O'Horo et al., 2015). The Brief Bedside Dysphagia Screening Test-Revised (BBDST-R) is a simple 8-item dysphagia screening tool with high sensitivity (95.5%; 95% confidence interval CI [CI] = 84.9–98.7%) and negative predictive value (88.9%; 95% CI = 67.2–96.9%) that is recommended for patients with neurological conditions (Mandysova et al., 2015). Both the preliminary version of the tool, the Brief Bedside Dysphagia Screening Test (BBDST) (Mandysova et al., 2011), and the BBDST-R (Mandysova et al., 2015) were developed in the Czech Republic based on a comparison of a simple bedside assessment and a quick swallow test with the gold standard, flexible endoscopic examination of swallowing. The largest subgroup was comprised of patients with stroke (54 of 112 patients or 48%) (Mandysova et al., 2015); therefore, the BBDST-R could be ideal for use in such patients.

Implicit in this approach is the assumption that nurses can agree on whether the result of the screening is normal or abnormal. Nonetheless, a recent systematic review of bedside dysphagia screening tests has found that reproducibility and consistency of such tests and protocols remain a challenge (O'Horo et al., 2015). Skills may vary between experienced andnovice nurses, which may affect decisions concerning the screening results. Therefore, determining the inter-rater reproducibility

Corresponding author: Petra Mandysová, Department of Nursing, Faculty of Health Studies, University of Pardubice, Průmyslová 395, Pardubice, Czech Republic, email: Petra.Mandysova@upce.cz

(agreement) of dysphagia screening tests used by nurses with varying levels of experience is important.

The BBDST-R, as a valid tool, has already been implemented in clinical practice (Mandysova et al., 2015). However, its inter-rater agreement has been reported only in a brief form (Mandysova, 2014), which is insufficient given the complexity of this measure. IRA is the degree to which item scores assigned by different raters are identical (Gisev et al., 2013). Obtaining high IRA is important not only from a theoretical but also from a practical point of view. If, at a given time, different nurses (raters) assess a particular patient using the tool, their results should ideally be identical. Because single summary measures of agreement (such as Cohen's kappa) provide only limited information, the authors of the Guidelines for Reporting Reliability and Agreement Studies (GRRAS) recommend reporting a combination of coefficients (measures) in order to obtain a more detailed impression of the degree of the agreement (Kottner et al., 2011).

Aim

The aim was to determine the inter-rater agreement (IRA) of the BBDST-R (Mandysova et al., 2015) for dysphagia screening in hospitalized patients with stroke, as assessed by nurses with varying levels of experience.

Methods

Design

This was a cross-sectional observational study of IRA of dysphagia screening performed and reported in accordance with the GRRAS (Kottner et al., 2011). The study involved nurse raters and patients hospitalized with stroke in the neurological department of a regional Czech hospital.

Sample

Two nurse raters were to conduct dysphagia screening in patients with stroke. As for the patients, convenience sampling was used. Enrolment could be performed only when both raters were immediately available in the department, which was approximately 2–3 hours once per week. Eligible patients were identified by nurses working in the department and they were approached by either of the two raters. The aim was to enrol, between August 2013 and December 2013 inclusive, at least 30 patients who would undergo a paired screening (one from Rater 1 and one from Rater 2). In a two-rater study, this sample size is sufficient to detect a kappa (κ) coefficient of 0.6 or higher as statistically significant ($p \leq 0.05$), with 90% power (more details concerning the κ coefficient are below), assuming the null hypothesis value of κ to be 0.00 (Sim, Wright, 2005).

The inclusion criteria were: clinically stable, sufficiently alert to understand and communicate intelligibly with the researcher, able to sit upright. The patients were enrolled only if they were willing to collaborate and sign an informed consent. Out of 51 patients who were approached, 5 refused to participate. Thus, the actual study involved 46 patients, 42 of whom had a paired screening: 28 (66.7%) men and 14 (33.3%) women (average age 72.4 ± 9.5; median age 73.5; age range 52–94).

Data collection

The Brief Bedside Dysphagia Screening Test-Revised (BBDST-R)

The raters were to conduct independent dysphagia screening using the BBDST-R (Mandysova et al., 2015). The BBDST-R consists of 8 bedside assessment items that are rated either as normal/negative (= 0) or abnormal/positive (= 1): a) "presence of voluntary cough", b) "ability to clench the teeth", c) "tongue symmetry and strength", d) "symmetry and strength of the facial muscles", e) "shoulder shrug symmetry and strength", f) "dysarthria", g) "aphasia", and h) "thick liquid: cough" (four teaspoons are given and the patient is observed for up to one minute afterwards) (Mandysova et al., 2015). The total score is obtained by summing up the scores of all 8 items and it is dichotomized (normal/"pass" versus abnormal/"fail") using a cut-off score of 1 (Mandysova et al., 2015). This means that the overall result is considered abnormal if at least one item is abnormal.

Rater Characteristics (Training and Experience)

Screening was conducted by two raters with varying levels of experience and different training. Rater 1 was a neuroscience nurse with over 20 years of clinical experience, with prior neuroscience training from a Canadian university-affiliated neurological hospital and neuroscience certification by the Canadian Nurses Association. She and her team developed the BBDST-R. Rater 2 was a novice nurse who was enrolled in a master-level nursing program and who had no prior clinical experience apart from her nursing studies. Rater 2 was trained by Rater 1. First, Rater 2 attended a 60-minute group training session led by Rater 1 (consisting of a 15-minute video, hands-on practice using simulation, and a discussion). Next, Rater 2 underwent 4-hour practical training in the neurological department where the study was to take place and practiced patient screening under the supervision of Rater 1.

The Procedure

For any given patient, the two independent assessments (one by Rater 1 and one by Rater 2) were conducted no more than 2 hours apart. The order in which the raters approached the patients depended mainly on their availability. The rater who assessed the patient first explained the purpose of the study to the patient and ensured that the informed consent was signed. Each rater was blinded to (did not witness) the other's assessment; however, both were aware that their judgments would be compared. The comparisons were performed only after all data had been collected. The assessments of Rater 1 were deemed the benchmark ("gold standard") against which the assessments of Rater 2 were compared.

Patients who did not undergo the 2nd screening (4 refused) were excluded from the analysis. Therefore, only 42 patients were included in the data analysis. For an analysis of the overall dichotomized result of the screening, only patients with paired assessments in all 8 items were included. For an analysis of individual items, only patients with paired assessments in that particular item were included; however, data could be missing in other items (e.g. patient refused or was unable to perform).

Data analysis

All data, i.e. individual item results as well as the overall, dichotomized result, were entered into a Microsoft Excel 2010 spreadsheet. Contingency tables with observed frequencies of normal / abnormal assessments by both raters were obtained using SPSS 20.0 statistical software (IBM SPSS, Inc., Chicago, Illinois). Cells *a* and *d* indicated, respectively, the number of patients for whom both raters agreed on the normal and abnormal result. Cells *b* and *c* indicated the number of patients on whom the raters disagreed; *n* = the number of patients with a paired assessment (once by each rater).

The IRA was measured using Cohen's kappa (κ). Kappa is commonly used for situations involving nominal variables and two raters and it corrects for agreement expected purely by chance (Sim, Wright, 2005; Gisev et al., 2013). For 2 × 2 contingency tables, κ can range from -1 to +1, where +1 indicates perfect agreement (Gisev et al., 2013). A value of 0 indicates exactly chance agreement, positive / negative values indicate that the observed agreement is greater / less than that expected from chance alone (Gisev et al., 2013). As for practical interpretation, a kappa value of ≤ 0.40 indicates poor agreement, 0.40–0.75 fair to good agreement, and ≥ 0.75 an excellent level of agreement (Gisev et al., 2013).

Kappa was computed with the mentioned SPSS statistical software, and the standard error of κ ($SE_{(κ)}$) was obtained for each value of κ. The significance level α = 0.01 or 0.05 (set by default in SPSS). Two-sided 95% confidence intervals (95% CI) were computed using a statistical calculator, *Kappa as a measure of concordance in categorical sorting* (Lowry, 2001–2013) or, for κ ≤ 0, using a Microsoft Excel 2010 spreadsheet, with the following formula: $κ - (1.96 \times SE_{(κ)})$ to $κ + (1.96 \times SE_{(κ)})$ (McHugh, 2012). However, the interpretation of κ is difficult as it is sensitive to the prevalence of the assessed attribute and imbalance (bias) between raters (Sim, Wright, 2005; Gisev et al., 2013). To overcome these challenges and aid the interpretation of κ, the following parameters were calculated from the contingency tables described above: the proportion of observed agreement (*Po*), using the formula $Po = (a + d) / n$; the prevalence index (p_{index}), using the formula $p_{index} = |a - d| / n$; and the bias index (b_{index}), using the formula $b_{index} = |b - c| / n$ (Sim, Wright, 2005).

Results

Of the 42 patients with paired assessments, 13 and 5 patients assessed by Rater 1 and 2, respectively, had some missing results. Data were missing mainly in item "thick liquid: cough" – in this item, 32 patients had a paired assessment (Table 1–2).

Table 1 Abnormal BBDST-R results by Rater 1 and Rater 2

BBDST-R Item	**n_1**	**n_2**	**n_1/n**
Overall dichotomized result (n† = 28)	24	17	0.86
Presence of voluntary cough (n = 40)	5	3	0.13
Ability to clench the teeth (n = 40)	1	7	0.03
Tongue symmetry and strength (n = 39)	17	10	0.44
Symmetry and strength of the facial muscles (n = 39)	22	10	0.56
Shoulder shrug symmetry and strength (n = 39)	12	14	0.31
Dysarthria (n = 40)	12	6	0.30
Aphasia (n = 40)	10	5	0.30
Thick liquid: cough (n = 32)	3	2	0.09

n – number of patients with a paired assessment in the analysed item (one assessment by Rater 1; one assessment by Rater 2); n† – number of patients with paired assessments in all items; n_1 – number of patients with an abnormal result by Rater 1; n_2 – number of patients with an abnormal result by Rater 2; n_1/n – number of patients with an abnormal result by Rater 1 (the "gold standard") expressed as the fraction of patients with a paired assessment

Table 2 Frequencies of BBDST-R assessment results, the proportion of observed agreement, the prevalence index, and the bias index

				Rater 1		**Total**	P_o	p_{index}	b_{index}
				Normal	Abnormal				
Overall dichotomized result (n = 28)	**Rater 2**	Normal	Fo	*a* = 3	*b* = 8	11	$\frac{3+16}{28} = 0.68$	$\frac{\vert 3-16\vert}{28} = 0.46$	$\frac{\vert 8-1\vert}{28} = 0.25$
		Abnormal	Fo	*c* = 1	*d* = 16	17			
	Total		Fo	4	24	*n* = 28			
				Rater 1		**Total**	P_o	p_{index}	b_{index}
				Normal	Abnormal				
Presence of voluntary cough (n = 40)	**Rater 2**	Normal	Fo	*a* = 35	*b* = 2	37	$\frac{35+3}{40} = 0.95$	$\frac{\vert 35-3\vert}{40} = 0.80$	$\frac{\vert 2-0\vert}{40} = 0.05$
		Abnormal	Fo	*c* = 0	*d* = 3	3			
	Total		Fo	35	5	*n* = 40			
				Rater 1		**Total**	P_o	p_{index}	b_{index}
				Normal	Abnormal				
Ability to clench the teeth (n = 40)	**Rater 2**	Normal	Fo	*a* = 32	*b* = 1	33	$\frac{32+0}{40} = 0.80$	$\frac{\vert 32-0\vert}{40} = 0.80$	$\frac{\vert 1-7\vert}{40} = 0.15$
		Abnormal	Fo	*c* = 7	*d* = 0	7			
	Total		Fo	39	1	*n* = 40			
				Rater 1		**Total**	P_o	p_{index}	b_{index}
				Normal	Abnormal				
Tongue symmetry and strength (n = 39)	**Rater 2**	Normal	Fo	*a* = 17	*b* = 12	29	$\frac{17+5}{39} = 0.56$	$\frac{\vert 17-5\vert}{39} = 0.31$	$\frac{\vert 12-5\vert}{39} = 0.18$
		Abnormal	Fo	*c* = 5	*d* = 5	10			
	Total		Fo	22	17	*n* = 39			
				Rater 1		**Total**	P_o	p_{index}	b_{index}
				Normal	Abnormal				
Symmetry and strength of the facial muscles (n = 39)	**Rater 2**	Normal	Fo	*a* = 16	*b* = 13	29	$\frac{16+9}{39} = 0.64$	$\frac{\vert 16-9\vert}{39} = 0.18$	$\frac{\vert 13-1\vert}{39} = 0.31$
		Abnormal	Fo	*c* = 1	*d* = 9	10			
	Total		Fo	17	22	*n* = 39			
				Rater 1		**Total**	P_o	p_{index}	b_{index}
				Normal	Abnormal				
Shoulder shrug symmetry and strength (n = 39)	**Rater 2**	Normal	Fo	*a* = 22	*b* = 3	25	$\frac{22+9}{39} = 0.79$	$\frac{\vert 22-9\vert}{39} = 0.33$	$\frac{\vert 3-5\vert}{39} = 0.05$
		Abnormal	Fo	*c* = 5	*d* = 9	14			
	Total		Fo	27	12	*n* = 39			
				Rater 1		**Total**	P_o	p_{index}	b_{index}
				Normal	Abnormal				
Dysarthria (n = 40)	**Rater 2**	Normal	Fo	*a* = 28	*b* = 6	34	$\frac{28+6}{40} = 0.85$	$\frac{\vert 28-6\vert}{40} = 0.55$	$\frac{\vert 6-0\vert}{40} = 0.15$
		Abnormal	Fo	*c* = 0	*d* = 6	6			
	Total		Fo	28	12	*n* = 40			
				Rater 1		**Total**	P_o	p_{index}	b_{index}
				Normal	Abnormal				
Aphasia (n = 40)	**Rater 2**	Normal	Fo	*a* = 30	*b* = 5	35	$\frac{30+5}{40} = 0.88$	$\frac{\vert 30-5\vert}{40} = 0.63$	$\frac{\vert 5-0\vert}{40} = 0.13$
		Abnormal	Fo	*c* = 0	*d* = 5	5			
	Total		Fo	30	10	*n* = 40			
				Rater 1		**Total**	P_o	p_{index}	b_{index}
				Normal	Abnormal				
Thick liquid: cough (n = 32)	**Rater 2**	Normal	Fo	*a* = 29	*b* = 1	30	$\frac{29+2}{32} = 0.97$	$\frac{\vert 29-2\vert}{32} = 0.84$	$\frac{\vert 1-0\vert}{32} = 0.03$
		Abnormal	Fo	*c* = 0	*d* = 2	2			
	Total		Fo	29	3	*n* = 32			

a – number of patients assessed as normal by both raters; b – number of patients assessed as abnormal by Rater 1 and normal by Rater 2; c – number of patients assessed as normal by Rater 1 and abnormal by Rater 2; d – number of patients assessed as abnormal by both raters; n – number of patients with a paired assessment; b_{index} *– bias index;* F_o *– observed frequency;* p_{index} *– prevalence index;* P_o *– proportion of observed agreement;* $P_o = (a + d) / n$; $p_{index} = |a - d| / n$; $b_{index} = |b - c| / n$

In all other individual items, paired assessments were achieved in 39–40 patients. Paired assessments in all items were achieved in 28 patients. Of these 28 patients, the overall dichotomized result was abnormal in 24 and 17 patients assessed by Rater 1 and 2, respectively.

Concerning the overall dichotomized result, both raters agreed in their assessments in 19 (68%) cases (3 normal and 16 abnormal); P_o = 0.68 (Table 2). As for the individual items, their agreement ranged from 31 (97%) cases for "thick liquid cough" to 22 (56%) cases for "tongue symmetry and strength". The p_{index} ranged from 0.18 ("symmetry and strength of the facial muscles") to 0.84 ("thick liquid: cough"), and it was 0.46 for the overall dichotomized result (Table 2). The b_{index} ranged from 0.03 ("thick liquid: cough") to 0.31 ("symmetry and strength of the facial muscles"), and it was 0.25 for the overall dichotomized result (Table 2).

Kappa ranged from -0.046 (p = 0.641) for "ability to clench the teeth" to 0.784 (p < 0.001) for "thick liquid: cough" (Mandysova, 2014) (Table 3). For the overall dichotomized result, κ was 0.241 (p = 0.114).

Table 3 Inter-rater agreement of the BBDST-R (Mandysova, 2014, p. 41)

BBDST-R Item	κ	p-value	SE(κ)	95% CI	
Overall dichotomized result (n† = 28)	0.241	0.114	0.161	0.000	0.558
Presence of voluntary cough (n = 40)	0.724	0.000**	0.183	0.366	1.000
Ability to clench the teeth (n = 40)	-0.046	0.641	0.041	-0.126	0.034
Tongue symmetry and strength (n = 39)	0.070	0.635	0.149	0.000	0.362
Symmetry and strength of the facial muscles (n = 39)	0.324	0.013*	0.119	0.091	0.557
Shoulder shrug symmetry and strength (n = 39)	0.540	0.000**	0.142	0.261	0.818
Dysarthria (n = 40)	0.583	0.000**	0.143	0.304	0.863
Aphasia (n = 40)	0.600	0.000**	0.153	0.299	0.900
Thick liquid: cough (n = 32)	0.784	0.000**	0.208	0.377	1.000

*n – number of patients with a paired assessment in the analysed item (one assessment by Rater 1; one assessment by Rater 2); n† – number of patients with paired assessments in all items; CI – confidence interval; κ – kappa; SE(κ) – standard error of kappa; * – κ is statistically significant p (≤ 0.05); ** – κ is statistically highly significant (p ≤ 0.01)*

Discussion

In health care research and practice, it is common that data is collected by multiple people; consequently, due to their variability, the question of agreement among them is relevant (McHugh, 2012). The study used two raters who represented the two extreme ends of the spectrum of clinical knowledge and experience: one was an expert and one was a novice. The second challenge was the fact that for most items of the BBDST-R (Mandysova et al., 2015), the two possible states (normal or abnormal) are not sharply differentiated. Consequently, the raters were required to make finer and even somewhat subjective discriminations while judging, for example, the strength of the tongue of the assessed patient against their own strength. Therefore, good agreement between the two raters was difficult to achieve. Similar challenges have been reported in other situations; for examples, in studies of pressure ulcers when variables included such items as amount of redness, oedema, and erosion in the affected area (McHugh, 2012).

In five BBDST-R items, the achieved IRA, as expressed by the P_o, was excellent (≥ 0.80) (Table 2). The P_o is easily interpretable; however, its key limitation is that it does not consider the possibility of agreement occurring purely by chance. Therefore, the P_o can overestimate the true IRA (Hugh, 2012). Hence, it is worthwhile to obtain κ, which is a chance-corrected measure.

The IRA of the overall dichotomized result and of three individual items ("ability to clench the teeth"; "tongue symmetry and strength"; "symmetry and strength of the facial muscles"), expressed by κ, was low (Table 3), reflecting the challenges mentioned above. On the other hand, four other items achieved a fairly good level of agreement, and one item ("thick liquid: cough") achieved an excellent level of agreement. Statistical significance was not always attained due to the lower sample size and low κ. It could have been achieved by planning the study with a bigger sample size, especially since missing data were likely to occur due to specific stroke-associated problems, such as aphasia and apraxia, which prevented some of the patients from better collaboration.

However, as already mentioned, the interpretation of κ is not straightforward because its magnitude can be affected by several factors. The major concern is that its behaviour is subject to prevalence of the assessed attribute (Shankar, Bangdiwala, 2014). If the prevalence is too low or too high, then the raters are more likely to agree either on the existence or non-existence of the attribute, which is reflected in a higher prevalence index (p_{index}) (Sim, Wright, 2005). At the same time, however, chance agreement is also higher, and paradoxically, κ is lower (Sim, Wright, 2005). This paradox occurs because κ represents agreement *beyond* chance (Sim, Wright, 2005).

The question therefore arises whether the prevalence of dysphagia in the sample was very high or very low. In research, true prevalence is not always known; instead, an experienced clinician's best judgment is used as a "gold standard" best estimate of the prevalence of the studied attribute (Cicchetti, 2011). In the present study, the results of Rater 1 were used for this purpose, and many or few abnormal results obtained by Rater 1, expressed as the fraction of patients with a paired assessment, were used to identify items in which the mentioned paradox may have been present (Table 1). Applying this approach, abnormality was rare in items "ability to clench the teeth", "thick liquid: cough" and "presence of voluntary cough", and it was very frequent in the overall dichotomized result.

Next, κ is affected by bias, i.e., the extent to which the raters disagree on the proportion of normal (or abnormal) cases (Sim, Wright, 2005). The paradox is that when there is a large bias (disagreement), κ is actually higher than in cases of low or absent bias (Sim, Wright, 2005). In our study, bias was expected due to the different levels of expertise of the raters.

To resolve the mentioned paradoxes, several authors recommend reporting κ in combination with other parameters, such as the prevalence index, and the bias index, and paying attention to situations where the p_{index} and the $\mathrm{b}_{\mathrm{index}}$ are high (Sim, Wright, 2005; Shankar, Bangdiwala, 2014). Other authors recommend reporting the original data in a contingency table alongside κ (Sim, Wright, 2005).

However, the use and reporting of multiple IRA parameters is frequently lacking (Shankar, Bangdiwala, 2014). For example, the authors of another dysphagia screening tool, the Gugging Swallowing Screen, reported the IRA using only two parameters: κ and the P_o (Trapl et al., 2007). Although the number of patients with a paired assessment was small (n = 20), the obtained κ was very high (0.773–1) and statistically highly significant ($p < 0.001$); the P_o was high as well (0.90–1.00) (Trapl et al., 2007). However, the b_{index} and the p_{index} or the original data were not provided, not allowing accurate interpretation of the obtained κ. Cichero et al. (2009) described the IRA of their dysphagia screening tool using solely a point estimate of κ, based on a very small sample size (n = 10).

As for our study, the p_{index} was high in all three rarely abnormal items (as per Rater 1): "ability to clench theteeth", "presence of voluntary cough" (0.80 in both), and "thick liquid: cough" (0.84) (Table 2), suggesting that the obtained κ in these items was actually an underestimation of the real IRA. As for the b_{index}, its highest value was obtained for item "symmetry and strength of the facial muscles" (Table 2), and in this particular case, the obtained κ was likely to be an overestimation of the real IRA based on Sim and Wright's (2005) advice on the b_{index} interpretation.

Despite the shortcomings of κ, it may assist researchers in uncovering discrepancies in diagnostic instruments. In fact, many studies have recently been carried out to measure the levels of agreement between experts and to identify factors that may contribute to the observed discrepancies so that instrument reliability could be improved (Nelson, Edwards, 2010). The factors that may influence agreement include the prevalence and severity of the condition, prior knowledge of the patient's clinical history and age, and the rater's level of experience and training (Nelson, Edwards, 2010). However, not all the factors can be eliminated. Instead, Hripcsak and Heitjan (2002) recommend separating the components of disagreement so that specific strategies (e.g., adjusting rater training) could be used to possibly improve agreement.

As for our study, determining the IRA of the BBDST-R (Mandysova et al., 2015) by means of statistical methods enabled us to uncover the fact that assessments by Rater 1 produced substantially more missing data than assessments by Rater 2, due to problems such as aphasia and apraxia. While this revelation does not reduce the validity of the BBDST-R itself, discrepancies among raters are not desirable and could be avoided by fine-tuning rater training and focusing on assessments of challenging patients. Next, to improve the IRA, it could be beneficial to differentiate the two possible states (normal or abnormal) more sharply by redefining some of the items (Martin, Bateson, 2007). For example, for item "thick liquid: cough", the raters could be taught that simple throat clearing does not represent cough. At present, the definitions of the BBDST-R items are not as detailed.

In summary, current clinical practice is placing an ever-increasing demand on patient assessments that are valid. To this end, various valid tools (instruments) are being implemented, meaning that they measure what they are purported to measure.

The BBDST-R (Mandysova et al., 2015) is one such tool. However, validity alone is not sufficient. The tool needs to produce results that are consistent. This consistency is of paramount importance to health care professionals including nurses because the results of the assessments guide decisions about further patient management and care. If, however, nurses do not assess the patient in a consistent way, then it is not clear whether apparent changes in the assessment results are due to actual changes in the patient's health status or due to some other factors, such as differences in assessment techniques. Striving to achieve the highest possible IRA is one way how to support consistency and thus contribute to correct clinical decision-making. Further research is needed to improve the IRA of the BBDST-R. At the same time, this study could spur new lines of nursing research inquiry that would focus on reliability of other new and existing measurement tools intended for clinical practice.

Conclusion

The study aimed to investigate the IRA of the 8-item BBDST-R (Mandysova et al., 2015) used in independent assessments of patients with stroke by two nurse raters. The study included an analysis of chance-corrected agreement and an exploration of factors that may have affected the results.

The IRA, as expressed by κ, was low for the overall result and variable for the individual items. Despite the fact that for some items, κ may have actually misrepresented the IRA due to such circumstances as a very low or very high prevalence of the condition and rater bias, several strategies were suggested to enhance the IRA of the instrument. The findings underscore the importance of studying and reporting reliability of instruments that are to be used in clinical practice.

Ethical aspects and conflict of interest

The study was conducted in accordance with ethical recommendations of the Helsinki Declaration of 1964, as revised in 2008, and as part of a larger multisite project, it was approved by the ethics committee of the hospital where the data collection took place as well as by the ethics committee of the principal investigator of the entire project. All participants were informed of the purpose of the study and agreed to be included in the research; they expressed this agreement by signing an informed consent form. Participation was voluntary, and all data were treated as confidential. The authors declare that the study has no conflict of interest.

Acknowledgements

The study was supported by IGA MZ NT 13725-4/2012

Author contribution

Conception and design (PM, EE, HT), data analysis and interpretation (PM), manuscript draft (PM), critical revision of the manuscript (PM, HT, EE), final version of the manuscript (PM).

References

Cicchetti DV. On the reliability and accuracy of the evaluative method for identifying evidence-based practices in autism. In: Reichow B, Doehring P, Cicchetti DV, Volkmar FR, editors. *Evidence-based practices and treatments for children with autism.* New York: Springer; 2011. p. 41–51.

Cichero JA, Heaton S, Bassett L. Triaging dysphagia: nurse screening for dysphagia in an acute hospital. *Journal of Clinical Nursing*. 2009;18(11):1649–1659.

Gisev N, Bell JS, Chen TF. Interrater agreement and interrater reliability: key concepts, approaches, and applications. *Research in Social & Administrative Pharmacy*. 2013;9(3):330–338.

Hripcsak G, Heitjan DF. Measuring agreement in medical informatics reliability studies. *Journal of Biomedical Informatics*. 2002;35(2):99–110.

Kottner J, Audige L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, Roberts C, Shoukri M, Streiner DL. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *International Journal of Nursing Studies*. 2011;48(6):661–671.

Lowry R. *Kappa as a measure of concordance in categorical sorting*; 2001–2013 [cited 2013 Dec 31]. Available from: http://vassarstats.net/kappa.html

Mandysova P. A vision for dysphagia screening by nurses. *Ošetrovateľstvo: teória, výskum, vzdelávanie*. 2014;4(1):37–41.

Mandysova P, Ehler E, Škvrňáková J, Černý M, Bártová I, Pellant A. Development of the Brief Bedside Dysphagia screening Test – Revised: a cross-sectional Czech study. *Acta Medica.* 2015;58(2):49–55.

Mandysova P, Škvrňáková J, Ehler E, Černý M. Development of the Brief Bedside Dysphagia Screening Test in the Czech Republic. *Nursing & Health Sciences*. 2011;13(4):388–395.

Martin P, Bateson P. *Measuring behaviour: An introductory guide.* 3 rd ed. Cambridge: Cambridge University Press; 2007.

Martino R, Martin RE, Black S. Dysphagia after stroke and its management. *Canadian Medical Association Journal.* 2012;184(10):1127–1128.

McHugh ML. Interrater reliability: the kappa statistic. *Biochemia Medica*. 2012;22(3):276–282.

Middleton S, Grimley R, Alexandrov AW. Triage, treatment, and transfer: evidence-based clinical practice recommendations and models of nursing care for the first 72 hours of admission to hospital for acute stroke. *Stroke*.

2015;46(2):e18–25. Nelson KP, Edwards D. Improving the reliability of diagnostic tests in population-based agreement studies. *Statistics in Medicine.* 2010;29(6):617–626.

O'Horo JC, Rogus-Pulia N, Garcia-Arguello L, Robbins J, Safdar N. Bedside diagnosis of dysphagia: a systematic review. *Journal of Hospital Medicine*. 2015;10(4):256–265.

Shankar V, Bangdiwala SI. Observer agreement paradoxes in 2x2 tables: comparison of agreement measures. *BMC Medical Research Methodology*. 2014;14:100.

Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical Therapy*. 2005;85(3):257–268.

Trapl M, Enderle P, Nowotny M, Teuschl Y, Matz K, Dachenhausen A, Brainin M. Dysphagia bedside screening for acute-stroke patients: the Gugging Swallowing Screen. *Stroke*. 2007;38(11):2948–2952.