JNM

Journal of Nursing Measurement

SPRINGER PUBLISHING COMPANY

www.springerpub.com/jnm

Journal of Nursing Measurement, Volume 25, Number 3, 2017

Inter-Rater Reliability and User-Friendliness of the Delirium Observation Screening Scale

Gerhard Mueller, Assoc.-Prof. Dr., RN
Petra Schumacher, MScN, RN
Jutta Wetzlmair, MScN, RN

UMIT – Private University for Health Sciences, Medical Informatics and Technology, Tyrol, Austria

Monika Lechleitner, Univ.-Prof. Dr., MD

State Hospital Hochzirl, Natters, Austria

Eva Schulc, Priv.-Doz. Dr.

UMIT – Private University for Health Sciences, Medical Informatics and Technology, Tyrol, Austria

Background and Purpose: Delirium is a common and often unrecognized complication of hospitalized elderly patients. Currently, there is no evidence for inter-rater reliability studies between registered nurses in the literature. Furthermore, the user-friendliness of the Delirium Observation Screening Scale (DOSS) has not been tested in Austria. Methods: A quantitative cross-sectional design with a convenience sample of 141 patients and 36 nurses in an Austrian hospital. Results: Analysis of rater-agreement and inter-rater reliability on item level, total score as well as category of delirium risk demonstrated very high agreement. In contrast, no or only fair kappa coefficient were determined. The user-friendliness of the scale was partially satisfactory. Conclusions: The very high absolute agreement speaks for the reliability of DOSS although the kappa paradox became obvious. The results of the presented study relate only to the tested setting.

Keywords: inter-rater reliability; user-friendliness; Delirium Observation Screening Scale (DOSS); acute care

Delirium is a frequent complication of hospitalized elderly patients older than 65 years and is caused by acute organic brain impairment with deterioration of cognitive abilities. Between 29% and 64% of older-than-65-year-old patients develop a delirium in the acute setting (Inouye, 2006; Inouye, Westendorp, & Saczynski, 2014; National Institute for Health and Clinical Excellence [NICE], 2010; Siddiqi, House, & Holmes, 2006). Symptoms fluctuate and may develop within a few hours to days. Because of the varying symptom presentation, delirium often stays undetected (Inouye et al., 2014; Schuurmans, Shortridge-Baggett, & Duursma, 2003; Siddiqi et al., 2006). Between 53% and 75% delirium cases remain unrecognized in hospitals (Mistarz, Eliott,

http://dx.doi.org/10.1891/1061-3749.25.3.504

Whitfield, & Ernest, 2011; Rice et al., 2011), 49%–87% in long-term care facilities (Voyer et al., 2012), and 46% in home care (Malenfant & Voyer, 2012). Delirium consequences are serious with high morbidity, prolonged hospital stay or even admission to a nursing home. Up to 40% of affected persons die as a result of delirium within a year (Schuurmans, Deschamps, Markham, Shortridge-Baggett, & Duursma, 2003). The differential diagnosis during hospital admission proves to be difficult because delirium and dementia symptoms may overlap (Fick, Hodo, Lawrence, & Inouye, 2007). In Austria, delirium is diagnosed according to the World Health Organization (WHO) International Statistical Classification of Diseases and Related Health Problems Criteria-10 (ICD-10). The severity of symptoms range from mild to severe and may cause emotional liability, changes to the sleep–wake cycle, psychomotor behavior, as well as cognitive impairment such as consciousness, attention, perception, thinking, and memory problems (WHO, 2016).

Standardized screening instruments may assist nurses in recognizing delirium early to initiate preventive interventions (Ijkema, Langelaan, van de Steeg, & Wagner, 2014). In recent years, many delirium screening instruments have been developed (Grover & Kate, 2012). A standardized instrument such as the Delirium Observation Screening Scale (DOSS) for successful delirium screening involves the situational, systematic, objective, and purposeful observation of a persons' behavior (Bierhoff & Petermann, 2014, p. 192).

BACKGROUND AND CONCEPTUAL FRAMEWORK

Nurses play a key role in recognizing temporary behavioral changes as well as changes in cognition and attention through their regular observation and continuity of patient care (Schuurmans, Deschamps, et al., 2003; Schuurmans, Shortridge-Baggett, et al., 2003). The concept of the perception process (Goldstein, 2015) is closely linked with patient observation. The dynamic and complex perception process consists of processing sensations of the environment in connection with ones' knowledge, which in turn leads to perceiving and recognizing a phenomena as well as acting based on the processed information (Goldstein, 2015, p. 3). The process of perception proceeds dynamically without a fixed order of the individual steps and can always change. Perception, recognition, and acting happen successively, simultaneously, as well as in reverse order and may affect each other (Goldstein, 2015, p. 4). The scientific observation is a situational, systematic, objective, and purposeful perception of phenomena, processes or a person's behavior and can thus be distinguished from an ordinary observation (Bierhoff & Petermann, 2014, p. 192). For a structured behavioral observation, a standardized observation system or monitoring instrument is used, which operationalize manifest variables of the observed occurrence. A standardized observation system can only be considered an accurate data collection instrument, if the measurement accuracy of each individual category is established with reasonable high observer or reliability coefficients (Döring & Bortz, 2016, p. 346)

Description, Administration, and Scoring of Delirium Observation Screening Scale

DOSS is a screening instrument for a possible delirium and initially consisted of 25 items, which were reduced to 13 verbal and nonverbal behavioral items (Schuurmans, Donders, Shortridge-Baggett, & Duursma, 2002). In each nursing shift, the 13 behavioral observations are rated in less than 5 min during regular nursing care with "never," "sometimes

TABLE 1. Items of the Delirium Observation Screening Scale

Items	The Patient
1	Dozes during conversation or activities
2	Is easily distracted by stimuli from the environment
3	Maintains attention to conversation or action
4	Does not finish question or answer
5	Gives answers which do not fit question
6	Reacts slowly to instructions
7	Thinks they are somewhere else
8	Knows which part of the day it is
9	Remembers recent events
10	Is picking, disorderly, restless
11	Pulls IV tubing, feeding tubes, catheters, and so forth
12	Is easily or suddenly emotional
13	Sees/hears things which are not there

Note. Rate items with "never" = 0 point, "sometimes to always" = 1 point. Reverse ratings of Items 3, 8, and 9. If unsure, answer "unable" = 0 points.

to always," and "unable." An item is rated "unable" if the behavior cannot be observed or recognized by the nurse because of knowledge deficits or patient's characteristics. The total value lies between 0 and 13 points with a cutoff value of 3 points and above indicating a possible delirium risk (Schuurmans et al., 2002). The German version of DOSS contains 13 items and was translated by Hasemann, Kressig, Ermini-Fünfschilling, Pretto, and Spirig (2007) based on scientific criteria. Table 1 presents the 13 items of the DOSS.

Procedures for Instrument Development

DOSS is based on the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed., *DSM-IV*) criteria (American Psychiatric Association, 1994) and was originally developed with geriatric medicine patients ($n = 82$) and elderly patients with hip fracture ($n = 92$) by Schuurmans, Shortridge-Baggett, et al. (2003). Of those patients, 4 geriatric medicine and 18 elderly patients with hip fracture were diagnosed delirious by a geriatrician. DOSS has a chosen cutoff value of 3 points (Schuurmans, Shortridge-Baggett, et al., 2003). For the 13-item DOSS version, the average score of three consecutive ratings during a 24-hr period demonstrated sensitivity of 94% and specificity of 78% (Schuurmans et al., 2002). The predictive validity of DOSS was determined against the *DSM-IV* criteria and reached sensitivity between 89% and 100% and specificity between 68% and 88% with a chosen cutoff value of 3 points (Schuurmans, Shortridge-Baggett, et al., 2003). In a palliative care setting ($n = 48$; Detroyer et al., 2014), sensitivity was 81.8% (95% confidence interval or CI [52, 95]) and specificity 96.1% (95% CI [90, 98]) determined by researchers with the Confusion Assessment Method (Inouye et al., 1990).

Internationally, DOSS was tested for content and construct validity (Schuurmans, Shortridge-Baggett, et al., 2003) as well as concurrent validity (Detroyer et al., 2014;

Scheffer, van Munster, Schuurmans, & de Rooij, 2011; Schuurmans, Shortridge-Baggett, et al., 2003). Content validity was determined by a group of seven experts (Schuurmans, Deschamps, et al., 2003). The construct validity was tested by correlations with the Informant Questionnaire of Cognitive Decline in Elderly (IQCODE; Jorm & Jacomb, 1989; $r = .33$ [$p < .05$] with $n = 82$ geriatric medicine patients; $r = .74$ [$p < .05$] with $n = 92$ patients with hip fracture) as well as the Barthel Index (Wade & Collin, 1988; $r = -.26$ [$p < .05$]; Schuurmans, Shortridge-Baggett, et al., 2003, $r = -.55$ [$p \leq .001$]). The concurrent validity between DOSS and the Delirium Index (McCusker, Cole, Bellavance, & Primeau, 1998) was tested with a palliative care sample ($n = 48$) and was moderate ($r_s = .53$, $p = .001$; Detroyer et al., 2014). Scheffer et al. (2011) determined concurrent validity ($r = .67$, $p = .01$) between the DOSS and Delirium Rating Scale-Revised-98 (DRS-R-98; Trzepacz et al., 2001) with 41 patients with hip fracture and 56 medical patients. Concurrent validity was also established by Schuurmans, Shortridge-Baggett, et al. (2003; $r = -.66$ [$p \leq .001$] with $n = 82$ geriatric medicine patients; $r = -.79$ [$p \leq .001$] with $n = 92$ patients with hip fracture) by correlating DOSS and the Mini Mental State Examination (MMSE; Folstein, Folstein, & McHugh, 1975).

The Cronbach's alphas were .93 ($n = 4$ delirious geriatric patients) and .96 ($n = 18$ delirious patients with hip fracture) and can be considered as acceptable indicators of internal reliability (Schuurmans, Shortridge-Baggett, et al., 2003). DOSS also showed internal consistency ($\alpha = .772$) in a palliative care setting (Detroyer et al., 2014).

User-friendliness of the DOSS was demonstrated by Detroyer et al. (2014) with 10 palliative care nurses and by van Gemert and Schuurmans (2007) with 39 medical and surgical nurses. Overall, nurses rated DOSS as easy to use, relevant for their daily work, as well as user-friendly (Detroyer et al., 2014; van Gemert & Schuurmans, 2007).

Objectives of Study

Currently, there is no evidence for inter-rater reliability studies between registered nurses in the scientific literature. Furthermore, the user-friendliness of DOSS has not been tested in Austria. Therefore, the aim of this study was to determine rater-agreement and inter-rater reliability of DOSS between two independent registered nurses in hospitalized medical patients and to ascertain the user-friendliness of DOSS from the perspective of registered nurses.

METHODS

Design and Sample

The study applied a quantitative-descriptive cross-sectional design. Data collection was carried out with two convenience samples at a hospital in Tyrol, Austria. All consecutive patients admitted to three units of the Department of Internal Medicine were included if they were at least 65 years of age and had an MMSE score higher than 20 points. Patients were excluded if they had a dementia diagnosis or were delirious on admission. The estimated sample size of 156 patients was based on a calculation formula by Gwet (2010). All registered nurses on the participating units were asked to take part in the study. All patients and 36 registered nurses were asked to give informed consent. The study was approved by the Ethics Committee of the Medical University of Innsbruck, Tyrol (Study identification number: AN2015-0268 355/4.17 356/4.1).

Procedure

This study was conducted between February 1 and May 9, 2016. Data collection began on admission date and continued on the second date. Prior to study beginning, participating registered nurses received a standardized hour-long training on study procedures and application of DOSS. In addition, nurses were informed to conduct the screenings within 1 hr and to not communicate their DOSS ratings to one another. Because of the fluctuating symptoms of delirium within days after admission (Schuurmans, Shortridge-Baggett, et al., 2003), screening was completed on admission date and on the second day. Two participating registered nurses on shift independently screened the patients with DOSS once in a 24-hr period based on their observations during routine patient care. Furthermore, patients' sociodemographic data were collected on admission date. The completed DOSS forms were collected in closed envelopes to ensure that ratings could not be compared. If a patient had a DOSS total score of 3 points or above, a physician was informed. A geriatric assessment including the MMSE was completed by a geriatrician on the third day following admission. Patients were excluded from data analysis retrospectively, if they had an MMSE score of 20 points or less.

At the end of the study, the participating 36 registered nurses received an adapted user-friendliness questionnaire. The questionnaire consisted of 17 user-friendliness items (Table 6) and 13 comprehensibility of single DOSS items (Table 7) with a comment section for each item. The 30 items were answered with a 5-point Likert scale (1 = *strongly agree*, 5 = *strongly disagree*). Those items were based on the studies by Detroyer et al. (2014) and van Gemert and Schuurmans (2007). The original authors did not discuss the psychometric qualities of the questionnaire. For this study, the questionnaire was supplemented with three open-ended questions on strength, weakness, and implementation of DOSS in nursing practice. In addition, one item was added about time to complete DOSS and two items on prior use of delirium screening scales. Nurses had 3 weeks to complete the questionnaire and were reminded daily by their head nurse to complete the questionnaire.

Statistical Analysis

Data analysis was performed with 141 patients' and 36 nurses' study records using SPSS Version 20.0. The sociodemographic data were analyzed descriptively on an exploratory level, depending on the level of measurement. In addition, percentage and absolute frequencies were calculated. The extend of agreement between the two registered nurses was measured with the overall percent agreement (Pa), chance-agreement probability (Pe), Cohen's kappa (κ), and intraclass correlation coefficient ($ICC_{1.1}$; Gwet, 2012; Wirtz & Caspar, 2002). Furthermore, the 95% CI was determined. To explore the strength of the association between the total scores of the DOSS, the Spearman's rho correlation coefficient (r_s) was used because the requirement for a parametric test was not fulfilled. The significance level was set at 5% ($\alpha \leq .05$).

The user-friendliness of DOSS was outlined with percentage and absolute frequencies as well as the median. Nurses' comments were systematized and grouped into categories.

RESULTS

Sample

One hundred forty-two patients signed informed consent, of whom 1 was excluded based on the MMSE score leaving 141 patients and 564 DOSS ratings (141 patients × 2 raters ×

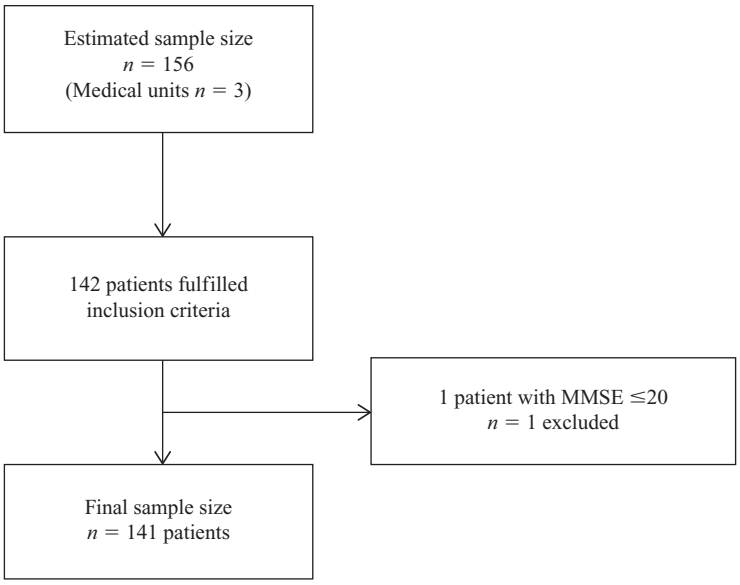

Figure 1. Patient sampling. Patients aged 65 years and older admitted to three units of the Department of Internal Medicine between February 1 and May 9 2016. MMSE = Mini Mental State Examination.

2 days) for data analysis. The estimated sample size of 156 patients could not be attained because of the short duration of the study. Figure 1 presents the sampling of the patients.

The patients had a mean age of 78.21 (*SD* ± 7.718) years, 69.1% were female, and 30.9% were male. The patients' mean MMSE score was 27.18 (*SD* ± 2.009). From the 36 participating registered nurses, 63.9% were female and 36.1% were male with an average age of 39.57 (*SD* ± 10.167) years. The registered nurses had on average 15.47 (*SD* ± 10.742) years of experience.

Occurrence Rates of Delirium Risk

A risk of delirium (agreement of both raters on total DOSS score ≥ 3 points) was present in 1 of the 141 patients (.71%) on the first assessment day. On the second assessment day, a risk of delirium was present in 2 of the 141 patients (1.42%).

Rater-Agreement and Inter-Rater Reliability

On item-level, high overall percent agreement was observed (Pa = 92.8%–100%) between the two independent registered nurses on the first assessment day. However, the calculated kappa coefficient showed no agreement ($\kappa = -.0017$ to .00) for 8 out of 13 items according to the rating by Landis and Koch (1977). This can be attributed to the high chance-agreement probability. Items 2 ($\kappa = .436$) and 10 ($\kappa = .562$) presented moderate agreement. Table 2 presents the rater-agreement on item-level for the first assessment day.

Table 3 presents the rater-agreement on item-level for the second assessment day.

High percent rater-agreement was also observed on the second assessment day (Pa = 92.7%–100%). Although, the calculated kappa coefficient showed poor to fair

TABLE 2. Rater-Agreement on Item-Level for Day 1

Items	*N*	Missing	Pa%	Pe%	κ	95% CI
1	141	0	98.6	98.6	−.007	[−.016, .003]
2	141	0	95.0	91.2	.436	[.089, .783]
3	137	4	94.2	94.3	−.0028	[−.049, −.006]
4	140	1	96.4	96.5	−.017	[−.032, −.001]
5	140	1	96.4	96.5	−.017	[−.032, −.001]
6	140	1	92.8	90.4	.255	[−.056, .567]
7	140	1	100	100	0	
8	139	2	95.7	94.4	.231	[−.170, .633]
9	137	4	95.6	94.3	.230	[−.171, .632]
10	140	1	97.8	95.1	.562	[.121, 1.003]
11	141	0	100	100	0	
12	141	0	97.9	97.9	0	
13	141	0	100	100	0	

Note. Pa% = overall percent agreement; Pe% = chance-agreement probability; CI = confidence interval.

TABLE 3. Rater-Agreement on Item-Level for Day 2

Items	*N*	Missing	Pa%	Pe%	κ	95% CI
1	140	1	100	100	0	
2	138	3	94.2	94.2	.304	[−.048, .657]
3	139	2	93.5	93.5	.148	[−.165, .462]
4	140	1	97.1	97.1	.320	[−.170, .462]
5	140	1	96.4	96.4	.268	[−.176, .713]
6	138	3	92.7	92.7	.248	[−.071, .567]
7	140	1	99.3	99.3	0	
8	139	2	95.7	95.7	−.022	[−.039, −.004]
9	139	2	95.0	95.0	−.019	[−.034, −.003]
10	139	2	97.1	97.1	−.009	[−.022, .005]
11	139	2	99.3	99.3	0	
12	140	1	99.3	99.3	0	
13	139	2	99.3	99.3	0	

Note. Pa% = overall percent agreement; Pe% = chance-agreement probability; CI = confidence interval.

TABLE 4. Rater-Agreement of Total Delirium Observation Screening Scale Scores for Day 1 and Day 2

	N	Missing	Pa%	Pe%	κ	95% CI	ICC	95% CI
Day 1	140	1	79.9	73.6	.214	[.065, .363]	.535	[.406, .643]
Day 2	140	1	78.5	73.8	.181	[.026, .336]	.383	[.233, .516]

Note. Pa% = overall percent agreement; Pe% = chance-agreement probability; CI = confidence interval; ICC = intraclass correlation coefficient.

agreement (Items 2–6) or no agreement at all because of high chance-agreement probability (Pe = 92.7%–100%).

For the DOSS total score, the overall percent agreement between the two independent registered nurses was 79.9% on the first and 78.5% on the second assessment day. The calculated chance-agreement probability values ranged between 73.6% and 73.8%. Table 4 presents the DOSS total scores for both assessment days.

For both assessment days, the calculated kappa coefficients showed poor to fair agreement (κ = .214 and κ = .181) and the ICC demonstrated moderate agreement (ICC = .535, 95% CI [.406, .643]) on the first assessment day and fair agreement (ICC = .383, 95% CI [.233, .516]) on the second assessment day. The determined correlation between the DOSS total scores was weak for both days (Day 1: $r_s = .364$, $p = .001$; Day 2: $r_s = .304$, $p = .001$). In addition, the overall percent agreement was determined between the two independent registered nurses for the category of a possible delirium risk (category <3 vs. ≥3 points) and is presented in Table 5.

The rater-agreement for the delirium risk category showed similar results that resemble those on item-level and for the total score. Again, the overall percent agreement between the two nurses was very high (Pa = 94.2% and 95%) for both days; although, the kappa coefficients showed poor agreement (κ = .170 and κ = .199) because of high chance-agreement probability values between 93.1% and 93.6%.

User-Friendliness of Delirium Observation Screening Scale

Out of 36 registered nurses, 27 completed the user-friendliness questionnaire of the DOSS (return rate 75%). None of the nurses had used DOSS prior to this study. On average, 11.19 (*SD* ± 7.254) ratings were completed by each nurse. The screening was done in less than 3 min by 18.5% ($n = 5$) of the nurses, 29.6% ($n = 8$) needed between 3 and 6 min, 11.1% ($n = 3$) needed 7–10 min, and six nurses (22.2%) took more than 10 min to complete the DOSS screening. Table 6 outlines the nurses' responses on the user-friendliness of DOSS, and Table 7 presents the comprehensibility of the single DOSS items.

TABLE 5. Rater-Agreement of Delirium Risk Category for Day 1 and Day 2

	N	Missing	Pa%	Pe%	κ	95% CI
Day 1	139	2	94.2	93.1	.170	[−.169, .509]
Day 2	140	1	95.0	93.6	.199	[.026, .336]

Note. Pa% = overall percent agreement; Pe% = chance-agreement probability; CI = confidence interval.

TABLE 6. User-Friendliness of Delirium Observation Screening Scale ($N = 27$)

Items	Strongly Agree[a]	Rather Agree	Partly Agree	Hardly Agree	Strongly Disagree	Mdn
Thorough delirium screening	3 (11.1)	9 (33.3)	13 (48.1)	1 (3.7)	1 (3.7)	3.00
Individual nursing assessment	1 (3.7)	10 (37.0)	9 (33.3)	2 (7.4)	5 (18.5)	3.00
Able to rate patient observation accurately	3 (11.1)	13 (48.1)	8 (29.6)	3 (11.1)		2.00
Concept of scale is clear	13 (48.1)	5 (18.5)	6 (22.2)	2 (7.4)	1 (3.7)	2.00
Language is compatible with practice	8 (29.6)	12 (44.4)	3 (11.1)	3 (11.1)	1 (3.7)	2.00
Observations are described free of values and judgment	15 (55.6)	9 (33.3)	3 (11.1)			1.00
Sufficient knowledge and experience to use scale	15 (55.6)	9 (33.3)	2 (7.4)		1 (3.7)	1.00
Observations can be rated differently	3 (11.1)	15 (55.6)	8 (29.6)	1 (3.7)		2.00
Clear differences between answer choices	5 (18.5)	9 (33.3)	7 (25.9)	6 (22.2)		2.00
Needed help from others, because of unclear task	5 (18.5)	3 (11.1)	7 (25.9)	2 (7.4)	10 (37.0)	3.00
Instructions on form were useful	7 (25.9)	9 (33.3)	10 (37.0)		1 (3.7)	2.00
Scale offers added value to practice	2 (7.4)	4 (14.8)	10 (37.0)	8 (29.6)	3 (11.1)	3.00
Scale assists in planning nursing care	1 (3.7)	5 (18.5)	10 (37.0)	7 (25.9)	4 (14.8)	3.00
Screening assists in implementing preventive nursing interventions	4 (14.8)	5 (18.5)	12 (44.4)	5 (18.5)	1 (3.7)	3.00
Scale serves as quality assurance tool ($n = 26$)	2 (7.4)	11 (40.7)	7 (25.9)	4 (14.8)	2 (7.4)	2.50
Scale is used in my line of work ($n = 26$)	5 (18.5)	3 (11.1)	4 (14.8)	6 (22.2)	8 (29.6)	4.00
Scale can be used as teaching material		10 (37.0)	10 (37.0)	3 (11.1)	4 (14.8)	3.00

Note. Mdn = median.
[a]*n* (%).

TABLE 7. Comprehensibility of Single Delirium Observation Screening Scale Items (n = 27)

Item	The Patient	Strongly Agree[a]	Rather Agree	Partly Agree	Hardly Agree	Strongly Disagree	Mdn
1	Dozes during conversation or activities	19 (70.4)	5 (18.5)	2 (7.4)	1 (3.7)		1.00
2	Is easily distracted by stimuli from the environment	20 (74.1)	5 (18.5)	2 (7.4)			1.00
3	Maintains attention to conversation or action	16 (59.3)	10 (37.0)	1 (3.7)			1.00
4	Does not finish question or answer	17 (63.0)	3 (11.1)	6 (22.2)	1 (3.7)		1.00
5	Gives answers which do not fit question	19 (70.4)	6 (22.2)	2 (7.4)			1.00
6	Reacts slowly to instructions	15 (55.6)	7 (25.9)	4 (14.8)		1 (3.7)	1.00
7	Thinks they are somewhere else	19 (70.4)	6 (22.2)	2 (7.4)			1.00
8	Knows which part of the day it is	18 (66.7)	7 (25.9)	2 (7.4)			1.00
9	Remembers recent events	16 (59.3)	8 (29.6)	3 (11.1)			1.00
10	Is picking, disorderly, restless	15 (55.6)	8 (29.6)	4 (14.8)			1.00
11	Pulls IV tubing, feeding tubes, catheters, and so forth	16 (59.3)	6 (22.2)	5 (18.5)			1.00
12	Is easily or suddenly emotional	14 (51.9)	8 (29.6)	5 (18.5)			1.00
13	Sees/hears things which are not there	16 (59.3)	6 (22.2)	5 (18.5)			1.00

Note. Mdn = median; IV = intravenous.
[a]*n* (%).

The registered nurses' comments on the strength of the DOSS concluded that the scale provides quick results in a short time period. In addition, the tool may support nurses in specific patient observation and in planning care. Using the instrument may sensitize nurses for possible delirium symptoms. Furthermore, DOSS is simple, clearly written, and easy to understand. However, the weaknesses and limitations of DOSS according to the nurses are shown in the incomprehensible wording and double negations of some items (Items 4, 6, 8, and 10). Two registered nurses thought that Item 13 is hard to assess on the first day of hospitalization. Nurses suggested adapting the scale linguistically to improve user-friendliness and comprehensibility of some items.

DISCUSSION

To our knowledge, this is the first study examining rater-agreement and inter-rater reliability of the DOSS between two independent registered nurses in a medical setting as well as to examining user-friendliness of DOSS from the perspective of medical nurses in an Austrian hospital. Screening tools may help nurses to identify the various aspects of delirium (Grover & Kate, 2012). Such instruments have to be reliable between different raters because they are used by nurses with different skill levels and experiences (Andrew et al., 2009). Therefore, the precision of the independent screening results between two registered nurses had to be investigated.

Generally, very high overall percent agreement was determined on both assessment days; however, the calculated kappa coefficients demonstrated poor or no agreement except for two items on the first assessment day. This study demonstrated that the calculation of Cohen's kappa can lead to unexpected results–high overall percent agreement with poor kappa coefficients. Therefore, the results suggest no agreement between registered nurses and are referred to as *kappa paradox* in the scientific literature (Feinstein & Cicchetti, 1990; Gwet, 2012, p. 36). For this purpose, Cohen's kappa considers the chance-agreement probability by calculating the raw and column marginal percentages of a contingency table (Grouven, Bender, Ziegler, & Lange, 2007). However, this consideration of the marginal percentages can cause the kappa paradox (Gwet, 2012, p. 37). The present results show high chance-agreement probabilities (Pe = 73.6%–100%). In addition, DOSS offers nurses only a limited number of response categories. This confinement increases the possibility of chance-agreement probability as well (Gwet, 2012, p. 6). Also, a homogeneous population may lead to poor kappa coefficients (Gwet, 2012, p. 38). The analyzed results are based on a homogeneous sample, whereas other studies included a diverse population of cognitively impaired elderly people with varying degrees of severity (Bhat & Rockwood, 2005). A different inter-rater reliability study of the NEECHAM Confusion Scale presented moderate agreement ($\kappa = .65$) between the raters (Neelon, Champagne, Carlson, & Funk, 1996) as compared to this study. Subsequently, the ICC was determined for the total DOSS score. The results present moderate agreement (ICC = .535) on the first assessment day and poor agreement (ICC = .383) on the second day. These results can be also attributed to a homogeneous population because variability of patient sample and rater-agreement must be given for high ICC values (Wirtz & Caspar, 2002, p. 161). The lower ICC on the second day might be caused by several factors. For example, delirium is a fluctuating phenomenon (Detroyer et al., 2014) and nurses might have taken interventions to prevent the risk of delirium or were less motivated to screen patients on the second day. The strength of the correlation between the total DOSS scores of the nurses was weak

for both assessment days (Days 1 and 2: $r_s = .364$ and $r_s = .304$). In contrast, the study by Schuurmans, Shortridge-Baggett, et al. (2003) presented moderate correlation ($r_s = .54$) between registered nurses and study nurses.

The results of the user-friendliness of DOSS were partially satisfactory. Almost half of the surveyed 27 nurses (48.1%) required less than 7 min for the use of DOSS. The other half contained either a concrete statement (18.5%) or took longer than 7 min (33%). One reason for the long assessment period may have been that the overall study recording was considered in answering the question. International studies (Schuurmans, Shortridge-Baggett, et al., 2003; Schuurmans et al., 2002) report less than 5 min for completing DOSS. Also, the user-friendliness of DOSS demonstrated inhomogeneous results. This is possibly because of the imminent implementation of the scale into the daily nursing routine. By contrast, most nurses were satisfied with the comprehensibility of the single DOSS items except of Items 4, 6, 8, and 10, which were vaguely formulated. Then again, DOSS was internationally proven to be very user-friendly (Detroyer et al., 2014; van Gemert & Schuurmans, 2007).

LIMITATIONS

One limitation of this study is that it was not the research interest to present the influence of raters' characteristics (e.g., age, experience) on the patients' ratings. Several possible biases may have influenced the results. The targeted number of at least 156 patients was not reached. This is mainly because of the short collection period. Therefore, there is the possibility that the power of the study was influenced by the reduced sample number. A selection bias is possible because of the convenience sampling strategy. This might have influenced the results of the questionnaire filled out by the registered nurses. An observer bias may be present because of nurses' observation during data collection. An interpretation bias might exist because of the interpretation of the DOSS ratings. Even though literature (Detroyer et al., 2014; van Gemert & Schuurmans, 2007) suggest to complete DOSS in three consecutive shifts in a 24-hr period to determine a mean value as indication of a risk for delirium, data collection was carried out only once during a 24-hr period. Because the registered nurses had only limited time resources, it was not possible for two nurses to rate patients consecutively 3 times in a 24-hr period. Another limitation is missing values on the DOSS. Possible reasons for missing ratings are workload of nurses, uncertainty in assessment of patient's behavior, nurses being fatigued or less motivated to participate in the study. Focused training and motivated nurses may reduce biases. However, training took place 3 months prior to study begin because of limited resources and delayed approval by the ethics committee resulting in a performance bias. Another limitation of the study was that a physician was only consulted if a patient had a total score of 3 or more points. Because of delirium symptoms may fluctuate during hospitalization, symptoms might have been missed by the attending physician or nurses.

RELEVANCE TO NURSING PRACTICE AND RESEARCH

In conclusion, the inter-rater reliability of the DOSS can be seen as acceptable because of high overall percent agreement, even though the kappa paradox was apparent for the tested setting. DOSS may assist nurses to structure their observation during regular patient care

and thus enables early identification of a possible delirium risk as well as early delirium management. Before implementing DOSS into the Austrian nursing practice, the Items 4, 6, 8, and 10 need to be linguistically adapted to make DOSS easier to understand. In addition, focused training is recommended for nurses on the use of DOSS. Further inter-rater reliability studies with a heterogeneous population are recommended for the clinical practice.

REFERENCES

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Andrew, M. K., Bhat, R., Clarke, B., Freter, S. H., Rockwood, M. R. H., & Rockwood, K. (2009). Inter-rater reliability of the DRS-R-98 in detecting delirium in frail elderly patients. *Age and Ageing*, *38*(2), 241–244.

Bhat, R., & Rockwood, K. (2005). Inter-rater reliability of delirium rating scales. *Neuroepidemiology*, *25*(1), 48–52.

Bierhoff, H. W., & Petermann, F. (2014). *Forschungsmethoden der Psychologie* [Research methods in psychology]. Göttingen, Germany: Hogrefe.

Detroyer, E., Clement, P. M., Baeten, N., Pennemans, M., Decruyenaere, M., Vandenberghe, J., . . . Milisen, K. (2014). Detection of delirium in palliative care unit patients: A prospective descriptive study of the Delirium Observation Screening Scale administered by bedside nurses. *Palliative Medicine*, *28*(1), 79–86.

Döring, N., & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial-und Humanwissenschaften* [Research methods and evaluation in the social and human sciences]. Berlin, Germany: Springer Verlag.

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, *43*(6), 543–549.

Fick, D. M., Hodo, D. M., Lawrence, F., & Inouye, S. K. (2007). Recognizing delirium superimposed on dementia: Assessing nurses' knowledge using case vignettes. *Journal of Gerontological Nursing*, *33*(2), 40–47.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state": A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, *12*(3), 189–198.

Goldstein, E. B. (2015). *Wahrnehmungspsychologie: Der Grundkurs* [Perception psychology: Basic course]. Berlin, Germany: Springer Verlag.

Grouven, U., Bender, R., Ziegler, A., & Lange, S. (2007). Der Kappa-Koeffizient [The kappa coefficient]. *Deutsche Medizinische Wochenschrift*, *132*(1), e65–e68.

Grover, S., & Kate, N. (2012). Assessment scales for delirium: A review. *World Journal of Psychiatry*, *2*(4), 58–70.

Gwet, K. L. (2010). *Inter-rater reliability: Sample size determination*. Retrieved from http://agreestat.com/blog_irr/sample_size_ determination.html

Gwet, K. L. (2012). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters* (3rd ed.). Gaithersburg, MD: Advanced Analytics.

Hasemann, W., Kressig, R. W., Ermini-Fünfschilling, D., Pretto, M., & Spirig, R. (2007). Screening, Assessment und Diagnostik von Delirien [Screening, assessment and diagnosis of delirium]. *Pflege*, *20*(4), 191–204.

Ijkema, R., Langelaan, M., van de Steeg, L., & Wagner, C. (2014). Do patient characteristics influence nursing adherence to a guideline for preventing delirium? *Journal of Nursing Scholarship*, *46*(3), 147–156.

Inouye, S. K. (2006). Delirium in older persons. *The New England Journal of Medicine*, *354*(11), 1157–1165.

Inouye, S. K., van Dyck, C. H., Alessi, C. A., Balkin, S., Siegal, A. P., & Horwitz, R. I. (1990). Clarifying confusion: The Confusion Assessment Method. A new method for detection of delirium. *Annals of Internal Medicine*, *113*(12), 941–948.

Inouye, S. K., Westendorp, R. G. J., & Saczynski, J. S. (2014). Delirium in elderly people. *Lancet*, *383*(9920), 911–922.

Jorm, A. F., & Jacomb, P. A. (1989). The Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE): Socio-demographic correlates, reliability, validity and some norms. *Psychological Medicine*, *19*(4), 1015–1022.

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.

Malenfant, P., & Voyer, P. (2012). Detecting delirium in older adults living at home. *Journal of Community Health Nursing*, *29*(2), 121–130.

McCusker, J., Cole, M., Bellavance, F., & Primeau, F. (1998). Reliability and validity of a new measure of severity of delirium. *International Psychogeriatrics*, *10*(4), 421–433.

Mistarz, R., Eliott, S., Whitfield, A., & Ernest, D. (2011). Bedside nurse-patient interactions do not reliably detect delirium: An observational study. *Australian Critical Care*, *24*(2), 126–132.

National Institute for Health and Clinical Excellence. (2010). *Delirium: Diagnosis, prevention and management*. Retrieved from http://www.nice.org.uk/guidance/cg103/resources/cg103-delirium-full-guideline3

Neelon, V. J., Champagne, M. T., Carlson, J. R., & Funk, S. G. (1996). The NEECHAM Confusion Scale: Construction, validation, and clinical testing. *Nursing Research*, *45*(6), 324–330.

Rice, K. L., Bennett, M., Gomez, M., Theall, K. P., Knight, M., & Foreman, M. D. (2011). Nurses' recognition of delirium in the hospitalized older adult. *Clinical Nurse Specialist*, *25*(6), 299–311.

Scheffer, A. C., van Munster, B. C., Schuurmans, M. J., & de Rooij, S. E. (2011). Assessing severity of delirium by the Delirium Observation Screening Scale. *International Journal of Geriatric Psychiatry*, *26*(3), 284–291.

Schuurmans, M. J., Deschamps, P. I., Markham, S. W., Shortridge-Baggett, L. M., & Duursma, S. A. (2003). The measurement of delirium: Review of scales. *Research and Theory for Nursing Practice*, *17*(3), 207–224.

Schuurmans, M. J., Donders, A. R., Shortridge-Baggett, L. M., & Duursma, S. A. (2002). Delirium case finding: Pilot testing of a new screening scale for nurses. *Journal of the American Geriatric Society*, *50*(4), S3.

Schuurmans, M. J., Shortridge-Baggett, L. M., & Duursma, S. A. (2003). The Delirium Observation Screening Scale: A screening instrument for delirium. *Research and Theory for Nursing Practice*, *17*(1), 31–50.

Siddiqi, N., House, A. O., & Holmes, J. D. (2006). Occurrence and outcome of delirium in medical in-patients: A systematic literature review. *Age and Ageing*, *35*(4), 350–364.

Trzepacz, P. T., Mittal, D., Torres, R., Kanary, K., Norton, J., & Jimerson, N. (2001). Validation of the Delirium Rating Scale-Revised-98: Comparison with the delirium rating scale and the cognitive test for delirium. *The Journal of Neuropsychiatry and Clinical Neurosciences*, *13*(2), 229–242.

van Gemert, L. A., & Schuurmans, M. J. (2007). The Neecham Confusion Scale and the Delirium Observation Screening Scale: Capacity to discriminate and ease of use in clinical practice. *BMC Nursing*, *6*, 3. http://dx.doi.org/10.1186/1472-6955-6-3

Voyer, P., Richard, S., McCusker, J., Cole, M. G., Monette, J., Champoux, N., . . . Belzile, E. (2012). Detection of delirium and its symptoms by nurses working in a long term care facility. *Journal of the American Medical Directors Association*, *13*(3), 264–271.

Wade, D. T., & Collin, C. (1988). The Barthel ADL Index: A standard measure of physical disability? *International Disability Studies*, *10*(2), 64–67.

World Health Organization. (2016). *ICD-10 Version: 2016. F05 Delirium, not induced by alcohol and other psychoactive substances*. Retrieved from http://apps.who.int/classifications/icd10/browse/2016/en#/F05

Wirtz, M., & Caspar, F. (2002). *Beurteileruebereinstimmung und Beurteilerreliabilitaet: Methoden zur Bestimmung und Verbesserung der Zuverlaessigkeit von Einschaetzungen mittels Kategoriensystemen und Ratingskalen* [Rater agreement and rater reliability: Methods for determining and improving the reliability of assessments using category systems and rating scales]. Göttingen, Germany: Hogrefe.

Acknowledgments. We thank the participating hospital, patients, and especially the registered nurses who supported us during data collection.

Correspondence regarding this article should be directed to Gerhard Mueller, Assoc.-Prof. Dr., RN, UMIT – Private University for Health Sciences, Medical Informatics and Technology, Department of Nursing Science and Gerontology, Eduard Wallnoefer-Zentrum 1, A-6060 Hall in Tyrol, Austria. E-mail: gerhard.mueller@umit.at