A randomized clinical trial using an educational intervention demonstrated no effect on interobserver agreement on assessments of functional status

by

Paquay L[1], De Lepeleire J[1], Milisen K[2,3], Ylieff M[4], Buntinx F[1,5]

Abstract

Aim

To evaluate the effect of an educational intervention on interobserver agreement of assessments of functional status performed by registered nurses and care assistants in a nursing home and to compare interobserver agreement in persons with and without cognitive impairment.

Background

High accuracy of assessments of functional status in care settings for older persons is needed for the efficacy of the planning and the evaluation of the nursing care.

Method

Randomized clinical trial. Six registered nurses and six care assistants were randomized to participate in an educational session about assessment instruments for functional status (intervention) or in a session about falls in the elderly (control). Each of the registered nurses and care assistants performed assessments on the same thirty-four residents using the Belgian Evaluation Scale (BES) and the AGGIR instrument. The kappa statistic (κ) for multiple observers (and its 95% confidence interval) was the main outcome measure.

Findings

At baseline, interobserver agreement for BES total score was: κ=0.43 (0.35-0.51) in the intervention group and κ=0.48 (0.39-0.57) in the control group. At the second assessment, agreement measures were: κ=0.48 (0.41-0.57) in the intervention group and κ=0.58 (0.50-0.66) in the control group. Results for AGGIR total scores were similar.

Conclusion

Interobserver agreement of assessments on nursing home residents was moderate and did not improve significantly after an educational session.

Keywords

Assessment, nursing, nursing home, reliability, randomized clinical trial, Belgian Evaluation Scale, Autonomie Gérontologie Groupe Iso-Ressources (AGGIR)

[1] Katholieke Universiteit Leuven, Department of General Practice, Leuven, Belgium

[2] Katholieke Universiteit Leuven, Centre for Health Services and Nursing Research, Leuven, Belgium

[3] University Hospitals of Leuven, Department of Geriatric Medicine, Leuven, Belgium

[4] Université de Liège, Faculty of Psychology and Educational Sciences, Liège, Belgium

[5] Universiteit Maastricht, Department of General Practice, Maastricht, The Netherlands
Correspondence: louis.paquay@med.kuleuven.be

Introduction

In nursing homes, assessment of functional status and other resident characteristics is used for different purposes, e.g. to determine resource utilisation, to provide a plan of care and to monitor outcome (1). In order to find out whether these assessments provide meaningful information, it is necessary to determine the reliability of the standardized instrument which is used by the professionals. The reliability of an assessment instrument is the degree of consistency with which it measures the attribute it is supposed to be measuring (2). Reliability is the concept one wishes to investigate when none of the multiple measurements is considered as ‘correct’ or as a ‘standard reference’. There is no criterion for the ‘correctness’ of judgements (3).

Many assessment instruments require that clinicians’ judgements are made into one of several mutual exclusive nominal or ordinal categories. Such an instrument is judged to be reliable if there is close agreement between multiple measurements. Defined as such, two types of reliability exist (4). First, when multiple clinicians independently use an assessment instrument for classifying the same subjects in discrete categories, the degree of agreement among the clinicians is an indicator of the interobserver reliability of the assessment instrument. Second, the degree of agreement between multiple assessments of a stable characteristic by the same observer is an indicator of intra-observer reliability. In the present study, the focus was on interobserver reliability.

The reliability of an instrument is linked to the population to which one wants to apply the instrument (5). Streiner and Norman stated that there is no such thing as the reliability of a test, unqualified. Reliability is relative and a reliability coefficient only has a meaning when applied to a specific population. For example, in a study using dual assessments of elderly nursing home residents by nurse assessors using the Health Care Financing Administration’s Minimum Data Set it was found that agreement concerning a resident’s activities of daily living status was significantly affected by a resident’s cognitive status (6). Assessments of residents suffering from cognitive impairment were significantly less reliable than assessments of cognitively intact residents.

Many authors recommend observer training to improve the reliability of a test (5). Cronin-Stubbs et al. (7) compared three programs for teaching nurses to use a functional assessment tool: simple training; training and practice using the Patient Evaluation and Conference System (PECS); training and collaboration with other nurses in group discussions. Training consisted of a one-hour lecture, a question and answer session, and a case study demonstration of a patient assessment. Training was associated with increased agreement among nurses in assessing functional status: the percent agreement, which was used as a measure for interrater agreement, was significantly higher ($p = 0.012$) in the treatment group of five nurses (median = 50%) than in the control group of five nurses (median = 20%). Adding practice or collaboration to the training session resulted in no differences between the treatment and the control group.

The effect of strict guidelines and a rigorous training program on variability in scoring the revised Acute Physiology and Chronic Health Evaluation (APACHE II) was investigated in 16 physicians (8). After implementation of a training program, interobserver agreement rates increased significantly from 59.7% to 76.5% and the interobserver reliability coefficient (weighted kappa) from 0.72 to 0.85.

In a multicenter international study of Alzheimer's disease, different initial training sessions resulted in improved interrater reliability, using the Alzheimer's Disease Assessment Scale (ADAS) for the assessment of videotapes of two older patients by 157 raters (9). Values of the intraclass correlation coefficient, which was the measure for interrater reliability used in the study, increased from 0.81 to 0.88 for patient A and from 0.91 to 0.97 for patient B. High values of the ICCs were maintained through refresher sessions in the course of the study.

The Belgian Evaluation Scale (BES) is used as a generic instrument for the assessment of functional status of older people with or without dementia and living at home or admitted to a care institution (10). The BES is an adaptation of the 'Index of ADL' (11).

Since its introduction in the early 1990s, interobserver agreement of BES assessments was a major problem: when the instrument was used for determining the dependency level of a resident and the corresponding remuneration of costs of nursing care, there often was disagreement between the advisor of the health insurance company and the registered nurse of the nursing home. In consequence, the Belgian National Institute for Sickness and Invalidity Insurance installed systematic control procedures on the consistent use of BES in homes for older persons and nursing homes. The procedures were recently updated (12).

The Autonomie Gérontologie Groupe Iso-Ressources scale (AGGIR) (13) was proposed by representatives of a major Belgian health insurance company as an alternative to the BES, although publications reporting high agreement on AGGIR assessments were not available at the time of the present study.

The main objective of the present study was to investigate whether an educational intervention resulted in higher interobserver agreement on assessments of functional status performed by registered nurses and care assistants in a nursing home, using the Belgian Evaluation Scale (BES) and the Autonomie Gérontologie Groupe Iso-Ressources scale (AGGIR). The second objective was to compare interobserver agreement in persons with and without cognitive impairment.

Methods

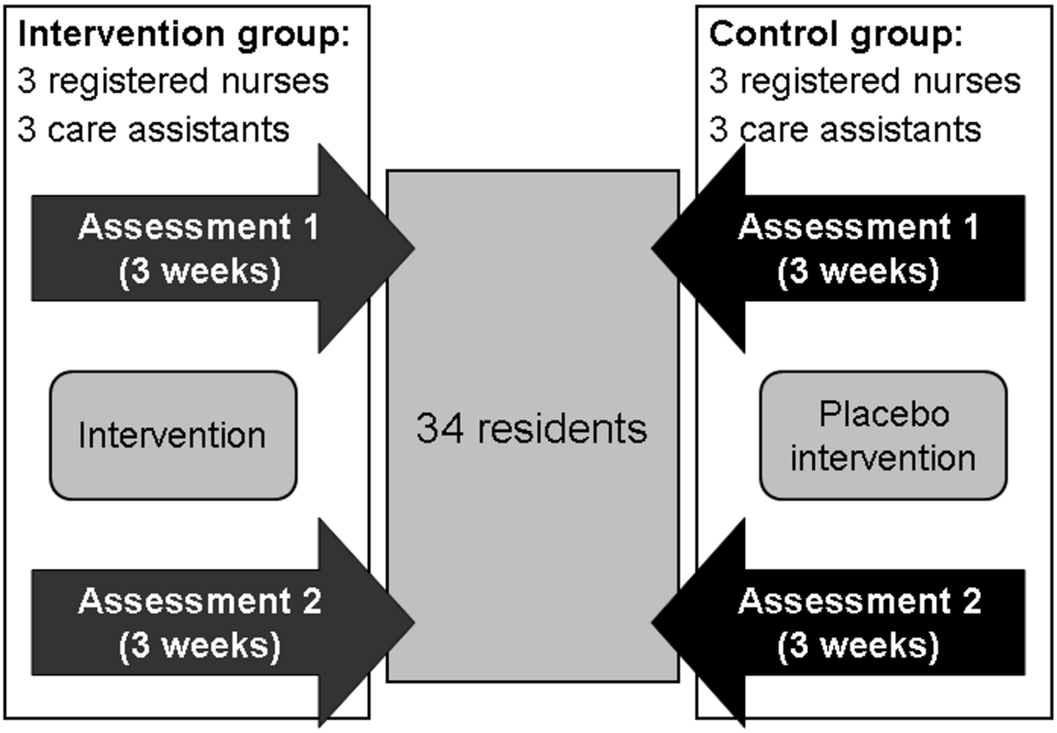
Sample

Thirty-four residents of a nursing home in a rural municipality in the province of Antwerp, Belgium, were assessed for functional status by six registered nurses and six care assistants, using the BES and the AGGIR scale. This specific small rural nursing home was selected from 40 nursing homes participating in the Qualidem study, because all nurses and care assistants were acquainted with the residents.

The residents were all eligible for the Qualidem study on assessment instruments for the care of older persons with dementia and were selected using a three-stage selection procedure. In the first stage, simple assessment of Activities of Daily Living (ADL), Instrumental Activities of Daily Living (IADL) (14) and behaviour were used to detect cognitive loss with great sensitivity. In the second and the third stage more specific diagnostic testing was performed among which the Mini-Mental State Examination (MMSE) (15) and finally the Cambridge Examination for Mental Disorders of the Elderly – Revised (CAMDEX-RN) (16) was used to select demented subjects. The selection procedure is described in detail elsewhere (17). The present study was carried out after the second stage. The residents were between 71 and 98 years of age (mean = 87 yrs, SD = 7 yrs); 25 (74%) were female. The residents' MMSE sum score ranged from 3 to 30; for 15 residents (44%) the MMSE was ≤ 23. All residents or their proxies provided written informed consent. The study used a protocol approved by the Ethical Committee of the University of Leuven Medical School.

Procedure

Figure 1. Study design



The registered nurses and care assistants of the intervention group and the control group received a manual with the instructions for the use of both instruments at the start of the first assessment period and performed the baseline assessments during three weeks before the intervention. The posttest assessments were done during a period of three weeks which started four days after the intervention. The delay between the baseline assessment and the posttest assessment could vary between 5 days and 6 weeks. The dates of the assessments were not registered. In the instructions for the assessors, it was stated explicitly that they were expected to perform the assessments individually and independently. It was chosen that both registered nurses and care assistants would do the assessments because health care workers of both professional disciplines are involved in carrying out functional assessments using the BES for several purposes in multiple care settings. General practitioners, nurses, care assistants and social workers have to use the instrument for interdisciplinary communication on an older person's functional status (18). In the present study, most nurses and care assistants had no previous training and had little experience in scoring both instruments.

The assessors were randomized to the intervention group and the control group, with three registered nurses and three care assistants in each group (Figure 1). The intervention consisted of a training session on guidelines for both instruments and a group discussion with medical advisors from health insurance organizations on correct interpretation of guidelines. During the training session and using the manuals of the BES and the AGGIR scale, the instructions for use of both scales were explained and the original video recording by the authors of the AGGIR scale was shown (19). As a placebo intervention, the control group participated in an educational session about the prevention of falls in the nursing home. The total time spent in the training session and the group discussion was two hours for both the intervention group and the controls.

Instruments

The Belgian Evaluation Scale (BES) (10) is an adaptation of the 'Index of ADL' (11). The instrument consists of six items which represent important activities of daily living: bathing, dressing, transferring, toileting, continence, and feeding. Each function has four (1=no assistance; 2=with assistive device or minimal assistance; 3=assistance; 4=total dependency) score categories. Two additional items on orientation in time and orientation in the living environment are scored from 1 to 5, with a higher score indicating a higher degree of disorientation. The total scale score can easily be deduced with the aid of a Boolean logic algorithm into one of five levels of dependency, which are coded with the capitals O (lowest dependency), A, B, C and Cd (highest dependency) (Table 1). In Belgium, the instrument is used for the evaluation of the functional status and the need for care in care institutions for older persons.

The Autonomie Gérontologie Groupe Iso-Ressources scale (AGGIR) is an assessment instrument for measuring the level of autonomy of older persons (13). The scale includes thirteen items (Table 3), which are coded A (full autonomy), B (partial autonomy) or C (no performance or full dependency). According to their item scores and with the help of a computer program, individuals are categorized in one of six levels of autonomy, with level 1 indicating the lowest and level 6 indicating the highest autonomy. In France, the instrument is used for determining care funding in institutions for older persons.

Table 1. Schematic representation of the Boolean logic algorithm used for classifying residents of Belgian nursing homes into five levels of dependency based on their item scores of the Belgian Evaluation Scale (BES) for the activities of daily living.

Items of the BES	Levels of dependency						
	O	A		B		C	Cd
		Physical dependency	Difficulties in orientation	Physical dependency	Difficulties in orientation	Physical dependency	Difficulties in orientation
Bathing	≤ 2	≥ 3	≤ 2	≥ 3	≥ 3	≥ 3	≥ 3
	AND	AND/OR	AND	AND	AND/OR	AND	AND
Dressing	≤ 2	≥ 3	≤ 2	≥ 3	≥ 3	≥ 3	≥ 3
	AND	AND	AND	AND	AND	AND	AND
Transferring	≤ 2	≤ 2	≤ 2	≥ 3	≤ 2	≥ 3	≥ 3
	AND	AND	AND	AND/OR	AND	AND	AND/OR
Toileting,	≤ 2	≤ 2	≤ 2	≥ 3	≤ 2	≥ 3	≥ 3
	AND	AND	AND	AND	AND	AND	AND
Continence	≤ 2	≤ 2	≤ 2	≤ 2	≤ 2	≥ 3	≥ 3
	AND	AND	AND	AND	AND	AND/OR	AND/OR
Feeding	≤ 2	≤ 2	≤ 2	≤ 2	≤ 2	≥ 3	≥ 3
	AND	AND	AND	AND	AND	AND	AND
Orientation in time	≤ 2	≤ 2	≥ 3	≤ 2	≥ 3	≤ 2	≥ 3
	AND	AND	AND	AND	AND	AND	AND
Orientation in the living environment	≤ 2	≤ 2	≥ 3	≤ 2	≥ 3	≤ 2	≥ 3

The Mini-Mental State Exam (MMSE) (15) is probably the most widely used screening measure of cognitive functioning. In the MMSE, different domains are assessed: orientation in time and place, registration of three words, attention and calculation, recall of three words, language, and visual construction. The maximum sum score is 30 points, indicating excellent cognitive function. A sum score of less than or equal to 23 was chosen as the cut-off for cognitive impairment (20).

Statistical analysis

The kappa statistic (κ) for multiple raters and its 95% confidence interval (CI) and the proportion of observed agreement were chosen as the measure for interrater agreement (21, 22). If there is complete agreement among the assessors, then $\kappa = 1$. If observed agreement is greater than or equal to chance expected agreement then $\kappa \geq 0$. If observed agreement is less than or equal to chance expected agreement, then $\kappa \leq 0$. The classification of Landis and Koch (23) was used for the interpretation of the relative strength of agreement associated with kappa statistics. The following labels were assigned to the corresponding ranges of κ: <0.00 poor; 0.00-0.20 slight; 0.21-0.40 fair; 0.41-0.60 moderate; 0.61-80 substantial; 0.81-1.00 almost perfect. We calculated the kappa statistic (and the 95% CI) with Microsoft® Excel software. Confidence intervals for kappas were calculated using the standard error. The extent of the agreement among the raters concerning each individual subject was calculated separately over all BES items and over all AGGIR items, using the formula of Siegel (21).

Paradoxes in κ values can be due to differences between two samples in the prevalence of an attribute (24-26). Two samples can have the same proportion of agreement on a condition between raters but if the prevalence of that condition is higher in one sample and almost all ratings will fall into one category, then κ will typically be lower. This paradoxical difference of the κ values arises because of the decision to impose a correction for chance agreement, making the assumption that the expected values for agreement should depend on the marginal totals. Since no assumptions are made about the marginal totals, two observers can get low values for κ despite a high percentage of observed agreement (24).

It must be emphasized that in the examples that are given in the publications mentioned above (24-26), the prevalence effect on the κ value demonstrated the effect of the imbalance of marginal totals of two response categories. In the present study, the items of the assessment instruments had three to five response categories. It is obvious that if there is an imbalance in the marginal totals of multiple response categories, then the κ value will also typically be lower.

In the present study, the proportion observed agreement is always presented next to the κ value in order to assess whether low κ value was due to low interobserver agreement or to the prevalence effect. If κ is low but the proportion agreement is high then it might be concluded that the measurement might to some extent be reliable. If κ is low and the proportion agreement between the assessors is low then the measurement is not reliable.

In order to test the hypothesis that agreement was most limited by a few outlier subjects, a sensitivity analysis was performed separately for each scale, excluding three residents with the lowest and the highest mean extent of agreement between raters on the item scores of each scale separately and recalculating the kappa estimates for total scale scores of the remaining 31 residents.

Graphical analysis was used for comparing interobserver reliability of subjects with and without cognitive impairment for each instrument and in each study condition separately: the pretest intervention group, the pretest control group, the posttest intervention and the posttest control group.

Results

Total scale scores

At baseline, all kappas referring to the agreement on total scale scores for BES and AGGIR indicated moderate agreement and were not significantly different between the intervention and the control group (Table 2). At the second assessment, all kappas referring to the agreement on total scale scores were higher than before the intervention, but the agreement was not significantly different between the second and the first assessment for BES and AGGIR total scale scores in both the intervention and the control condition.

Table 2. Kappa (κ) and its 95% confidence limits and the proportion observed agreement as measures of agreement between multiple raters about Belgian Evaluation Scale assessments and AGGIR assessments

Assessment instrument	Study condition	First assessment		Second assessment	
		κ (95% CI)	Proportion observed agreement	κ (95% CI)	Proportion observed agreement
BES	Intervention group	0.43 (0.35-0.51)	0.61	0.48 (0.41-0.57)	0.65
	Control group	0.48 (0.39-0.57)	0.65	0.58 (0.50-0.66)	0.71
AGGIR	Intervention group	0.55 (0.50-0.60)	0.61	0.58 (0.54-0.63)	0.64
	Control group	0.53 (0.49-0.58)	0.59	0.54 (0.50-0.58)	0.59

The number of assessors was 6 per study condition and the number of assessed nursing home residents was 34.

Item scores

At baseline, agreement on BES and AGGIR item scores was not different between the intervention and the control group: e.g. for BES washing, κ = 0.43 (95% CI 0.37-0.49) in the intervention group and κ = 0.44 (95% CI 0.38-0.50) in the control group (Table 3). At the second assessment, only the kappa for the BES item washing (κ = 0.63 [95% CI 0.56-0.70]), was significantly higher than the kappa for washing of the first assessment (κ = 0.44 [95% CI 0.38-0.50]). All other item kappas were not significantly different between two assessments, for either the intervention group or the control group (Table 3).

Sensitivity analysis

After excluding the three residents with the lowest or the highest mean extent of agreement per instrument from the analysis, kappas referring to the agreement on total scale scores were higher and lower respectively, and were not significantly different between the intervention and the control group (data not shown in this paper). All other trends were similar to trends for the agreement in the total group.

Graphical analysis

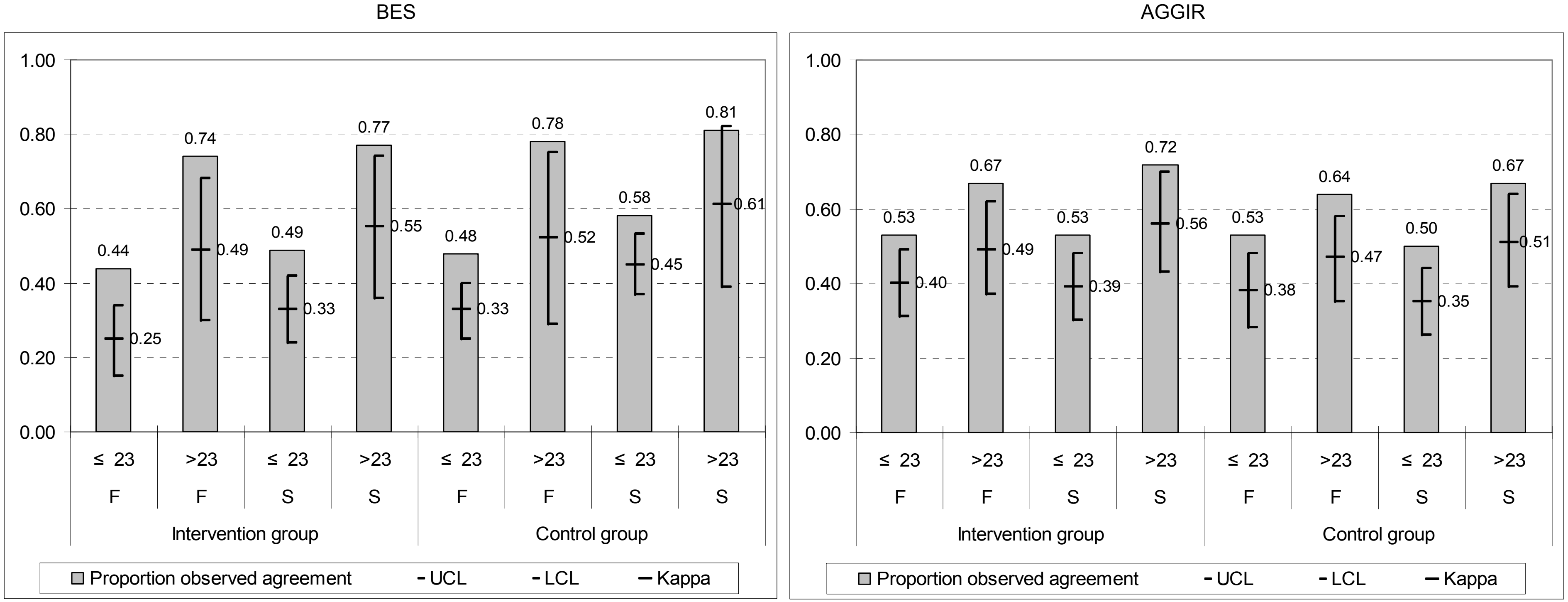
In Figure 2, the proportions observed agreement (vertical bars) and kappa values with its 95% confidence intervals are represented in adjacent graphs for residents with cognitive impairment (≤ 23) and without cognitive impairment (> 23) and for each study condition (intervention group and control group; first and second assessment). Kappa values are shown beside the confidence intervals, the value of the proportions agreement are shown on top of the vertical bars. Although there is no significant difference between kappas of residents with and without cognitive impairment within any of the study conditions, a consistent pattern is apparent: in each study both the proportion agreement and the kappa value is higher for residents without cognitive impairment than for residents with cognitive impairment.

Table 3. Kappa (κ) and its 95% confidence interval and the proportion observed agreement as measures of agreement between multiple raters about assessments using the Belgian Evaluation Scale (BES) and the AGGIR scale, before and after an educational intervention in a nursing home.

Instrument and items		Intervention group				Control group			
		First assessment		Second assessment		First assessment		Second assessment	
		K (95% CI)	P_o	K (95% CI)	P_o	K (95% CI)	P_o	K (95%CI)	P_o
BES	Washing	0.43 (0.37-0.49)	0.58	0.48 (0.42-0.54)	0.63	0.44 (0.38-0.50)	0.60	0.63 (0.56-0.70)	0.74
	Clothing	0.45 (0.39-0.50)	0.59	0.53 (0.47-0.60)	0.67	0.48 (0.41-0.54)	0.62	0.56 (0.48-0.63)	0.71
	Transfer	0.62 (0.55-0.68)	0.73	0.51 (0.44-0.57)	0.65	0.55 (0.48-0.61)	0.68	0.58 (0.52-0.64)	0.70
	Toileting	0.56 (0.43-0.69)	0.76	0.51 (0.39-0.62)	0.72	0.48 (0.36-0.59)	0.69	0.53 (0.42-0.65)	0.73
	Continence	0.29 (0.19-0.39)	0.56	0.27 (0.18-0.36)	0.54	0.40 (0.28-0.52)	0.66	0.39 (0.28-0.50)	0.63
	Eating	0.44 (0.29-0.60)	0.75	0.37 (0.24-0.50)	0.68	0.27 (0.10-0.43)	0.66	0.37 (0.19-0.54)	0.74
	Orientation in time	0.43 (0.31-0.55)	0.66	0.48 (0.38-0.59)	0.69	0.43 (0.31-0.54)	0.66	0.54 (0.43-0.66)	0.73
	Oriëntation in place	0.43 (0.28-0.58)	0.71	0.42 (0.29-0.55)	0.68	0.38 (0.23-0.53)	0.69	0.49 (0.35-0.63)	0.74
AGGIR	Coherent behaviour	0.47 (0.35-0.60)	0.73	0.49 (0.38-0.59)	0.73	0.45 (0.33-0.58)	0.72	0.33 (0.23-0.44)	0.63
	Orientation	0.57 (0.43-0.70)	0.78	0.48 (0.37-0.59)	0.71	0.48 (0.35-0.61)	0.72	0.52 (0.38-0.65)	0.75
	Personal hygiene of upper body parts	0.47 (0.38-0.55)	0.67	0.53 (0.45-0.60)	0.69	0.56 (0.49-0.64)	0.72	0.60 (0.52-0.67)	0.75
	Personal hygiene of lower body parts	0.54 (0.47-0.61)	0.70	0.61 (0.54-0.68)	0.75	0.59 (0.51-0.66)	0.73	0.57 (0.50-0.65)	0.73
	Clothing: upper	0.51 (0.43-0.58)	0.69	0.60 (0.53-0.68)	0.74	0.60 (0.52-0.67)	0.74	0.59 (0.51-0.66)	0.74
	Clothing: middle	0.48 (0.41-0.56)	0.67	0.61 (0.53-0.68)	0.75	0.54 (0.46-0.62)	0.71	0.67 (0.59-0.75)	0.80
	Clothing: under (underwear)	0.51 (0.44-0.57)	0.67	0.63 (0.56-0.70)	0.75	0.55 (0.48-0.62)	0.71	0.63 (0.56-0.71)	0.77
	Feeding: serve food	0.45 (0.33-0.57)	0.72	0.31 (0.20-0.42)	0.62	0.36 (0.25-0.47)	0.64	0.40 (0.28-0.51)	0.66
	Feeding: eating	0.33 (0.05-0.61)	0.82	0.25 (-0.10-0.60)	0.85	0.21 (0.05-0.38)	0.66	0.29 (0.02-0.56)	0.79
	Urinary elimination	0.46 (0.32-0.60)	0.73	0.51 (0.37-0.65)	0.75	0.58 (0.46-0.70)	0.77	0.49 (0.35-0.62)	0.74
	Faecal elimination	0.43 (0.28-0.59)	0.73	0.52 (0.38-0.66)	0.75	0.53 (0.40-0.65)	0.74	0.47 (0.34-0.60)	0.72
	Transfers (get up, go to bed, …)	0.55 (0.43-0.66)	0.74	0.58 (0.48-0.68)	0.75	0.43 (0.34-0.53)	0.66	0.50 (0.39-0.60)	0.71
	Ambulation inside the house	0.53 (0.45-0.62)	0.71	0.54 (0.46-0.62)	0.71	0.49 (0.41-0.57)	0.68	0.49 (0.41-0.58)	0.68

The number of assessors was 6 per study condition and the number of assessed residents was 34.
CI = confidence interval
P_o = proportion observed agreement

Figure 2. Kappas (95% confidence limits) and proportions observed agreement as measure of agreement between multiple raters about assessments using the Belgian Evaluation Scale (BES) and the AGGIR scale, before and after a randomized controlled educational intervention in a nursing home



The number of assessors was 6 per study condition and the number of assessed nursing home residents was 15 for the group with cognitive impairment (MMSE sum score ≤ 23) and 19 for the group who were cognitively intact (MMSE sum score > 23).

≤ 23 = MMSE sum score ≤ 23; > 23 = MMSE sum score > 23

F = first assessment; S = second assessment

UCL = upper 95% confidence limit; LCL = lower 95% confidence limit

Discussion and conclusions

In this study, interobserver agreement of assessments on nursing home residents was moderate and did not improve significantly after an educational session. At the second assessment, all kappas referring to total scale scores were higher than the corresponding kappas of the first assessments. Although most registered nurses and care assistants had no previous training and had little experience in scoring both instruments, an interventional training session did not influence the aptitude of the intervention group significantly. Several factors might have contributed to the slight increase of the agreement on functional assessments in both study groups: recall effects; the Hawthorne effect; nurses and care assistants may have got used to assess residents; maybe they were motivated by the attention of the researchers to perform assessments with higher accuracy.

Based on this conclusion, it seems important for practice not to expect too much effect of a single educational session for improving interobserver agreement on functional assessments. It might rather be recommended to make use of a repeated and multifaceted strategy which emphasizes on discussion and mutual consultation between the assessors about the interpretation of the instructions for use of the assessment instrument. In the present study, the instructions to the registered nurses and care assistants explicitly stated not to discuss the residents' functioning at the time when they were carrying out the assessments. The effectiveness of the educational session was limited because the assessors of the intervention group had only limited occasion for discussion, which may not have been sufficient to clarify dissenting interpretations of score categories. The fact that assessments of the same residents may have been performed on different dates might also have affected interobserver reliability, but these effects could not be accounted for because the dates of the assessments were not registered.

Another limitation of the study population was that the sample may have been too small to yield significant differences between two kappa values. The statistical power of the present study may also have been reduced by skewed score distributions. However, there are some indications that κ estimates in the present study were robust. First, when comparing agreement measures for both assessment instruments or for both study conditions, trends were very similar: e.g. with regard to interobserver agreement on single category assignment of the total scale score, agreement was highest in the categories indicating lowest dependency (according to BES) and highest autonomy (according to AGGIR). Secondly, recalculating kappas excluding three residents with the lowest agreement on the items of each scale yielded confidence intervals which were not much larger than the confidence intervals of kappas for agreement on total group scores. Thirdly, the dependence of κ on the observed marginal prevalences seems limited in this study. The difference between the proportion observed agreement and κ was highest with regard to the AGGIR item feeding-eating: e.g. at baseline the proportion observed agreement in the intervention group was 0.82 and κ=0.33; the relative distribution of feeding-eating scores showed a major imbalance in marginal totals: 84% of all 204 scores by 6 observers on 34 residents were in category A (n=172); 14% in category B (n=29); 2% in category C (n=3).

With regard to the second objective, the comparison of interrater agreement between persons with and without cognitive impairment, a consistent pattern was demonstrated over all study conditions and for both scales: observed agreement and kappas referring to persons without cognitive impairment were consistently higher. Although there was no statistical significant difference, the repeated pattern of these findings might be interpreted as an indication that assessments of residents suffering from cognitive impairment were less reliable than assessments of cognitively intact residents, which might be a confirmation of earlier findings (6). Probably, both assessment instruments were insufficiently adapted for taking into account the specific characteristics of functional performance associated with cognitive impairment. In fact, these instruments were originally intended for general use in a population of older persons and not specifically for use in persons suffering from cognitive impairment and dementia. Specific assessment instruments may be more adequate for the assessment of cognitively impaired persons. For example, the Abilities Assessment Instrument (AAI) was developed to assess the self-care, social, interactional and interpretive abilities of older people with cognitive impairment related to dementia (27). Another alternative might be the Bedford Alzheimer Nursing Severity Scale for the Severely Demented, which combines ratings of cognitive (speech, eye contact) and functional deficits (dressing, eating, ambulation) with occurrence of pathological symptoms (sleep-wake cycle disturbance, muscle rigidity/contractures) (28, 29). An aspect that might have caused lower interrater agreement in the present study might have been the tendency of persons suffering from cognitive impairment or dementia to hide their functional deficits and to present, at least partially, a facade of normal functional performance (30). This tendency may have confused the assessors and caused higher disagreement on the functional status of residents with cognitive impairment.

In a future study, it seems appropriate to compare the reliability (and validity) of a specific instrument for assessment of persons with cognitive impairment with general assessment instruments such as the BES or AGGIR.

References

1. Morrissey Y, Bowman C & Carpenter I. Assessment of patients in long-term care should be used to improve quality as well as allocate funds. Editorial. Age and Ageing 2006; 35: 212-4
2. Polit DF, Hungler BP. Essentials of nursing research. Philadelphia, J.B. Lippincott Company, 1995
3. Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 1960; Vol. XX: 37-46
4. Sim J, Wright CC. The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. Physical Therapy 2005; 85: 257-68
5. Streiner DL, Norman GR. Health measurement scales. A practical guide to their development and use. Third edition. Oxford, Oxford University Press, 2003
6. Phillips CD, Chu CW, Morris JN, Hawes C. Effects of Cognitive Impairment on the Reliability of Geriatric Assessments in Nursing Homes. Journal of the American Geriatrics Society 1993; 41: 136-42

7. Cronin-Stubbs D, Swanson B, Dean-Baar S, Sheldon JA, Duchene P. The Effects of a Training Program on Nurses' Functional Performance Assessments. Applied Nursing Research 1992; 5(1): 38-43
8. Polderman, KH, Jorna EMF, Girbes ARJ. Inter-observer variability in APACHE II scoring: effect of strict guidelines and training. Intensive Care Medicine 2001; 27: 1365-9
9. Yekrangi-Hartmann C, Bernhardt T, Baltissen R. Ratertraining to Establish Interrater Reliability in a Study with Alzheimer Patients [Trainingsmassnahmen zur Verbesserung der Interrater-Reliabilität in einer Alzheimer-Studie]. Zeitschrift für Gerontopsychologie & Geriatrie 1999; 12(3): 143-55
10. Arnaert A, Delesie L. Calibration of measurement data: the Belgian Institute of Health Insurance Index of ADL. [IJking van meetgegevens: RIZIV A.D.L.-index]. Acta Hospitalia 1999; 39(4): 19-31
11. Katz S, Ford A., Moskowitz R, Jackson B, Jaffe M. Studies of Illness in the Aged. JAMA 1963; 185: 914-9
12. Rijksinstituut voor Ziekte- en Invaliditeitsverzekering / Institut National d'Assurance Maladie-Invalidité (RIZIV/INAMI). Royal Decree of August 21, 2008 and information report on control procedures of the Belgian Evaluation Scale. [Koninklijk besluit van 21 augustus 2008 over de nieuwe "Kappa"-controles in de ROB's en RVT's. Nieuwe Brochure over de "Kappa"-controles in de ROB's en RVT's (vanaf 1 september 2008)] Website documents, available on: http://www.riziv.fgov.be/care/nl/residential-care/specific-information/katz.htm (Website consulted February 28, 2009)
13. Vetel J-M, Leroux R, Ducoudray J-M, Prévost P. AGGIR, Precisions on it's origin, practical instructions for use. [AGGIR, Précisions sur sa genèse, Conseils pratiques d'utilisation]. La Revue du Généraliste et de la Gérontologie 1998; 47: 27-33
14. Lawton MP, Brody EM. Assessment of older people: Self-maintaining and instrumental activities of daily living. Gerontologist 1969; 9 (3): 179-86
15. Folstein MF, Folstein SE, McHugh PR. 'Mini-Mental State': a practical method for grading the cognitive state of patients for the clinician. Journal of Psychiatric Research 1975; 12: 189-98
16. Roth, M., Tym, E., Mountjoy, C.Q., Huppert, F.A., Hendrie, H., Verma, S., Goddard, R., 1986. CAMDEX – A standard instrument for the diagnosis of mental disorder in the elderly with special reference to the early detection of dementia. British Journal of Psychiatry 149, 698-709
17. Paquay L, De Lepeleire J, Schoenmakers B, Stessens J, Bouwen A, van der Burg M, Di Notte D, Gazon R, Ylieff M, Fontaine O, Buntinx F. The Qualidem project in Belgium. A two-center study on care needs and provision in dementia care: Inclusion criteria and description of the population. Archives of Public Health 2004; 62: 145-62
18. De Lepeleire J, Paquay L, Jacobs M. De verschillende schalen voor ADL-activiteiten voor volwassenen in de Vlaamse gezondheidszorg. Huisarts Nu 2005; 34 (2): 58-68
19. Leroux R, Vetel J-M, Ducoudray J-M. AGGIR. Videotape. France, Syndicat National de Gérontologie Clinique, 1999
20. Tombaugh TN, McIntyre NJ. The Mini Mental State Examination: A comprehensive review. Journal of the American Geriatrics Society 1992; 40: 922-35
21. Siegel S, Castellan NJ. Nonparametric statistics for the Behavioral Sciences. New York, Mc Graw-Hill, 1988
22. Fleiss JL, Levin B, Paik MC. Statistical Methods for Rates and Proportions. Third Edition. Hoboken NJ, John Wiley & Sons Inc., 2003
23. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33: 159-74
24. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. Journal of Clinical Epidemiology 1990; 43: 543-9
25. Hoehler FK. Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. Journal of Clinical Epidemiology 2000; 53: 499-503
26. Tooth LR, Ottenbacher KJ. The κ Statistic in Rehabilitation Research: An Examination. Archives of Physical and Medical Rehabilitation 2004; 85: 1371-6

27. Dawson P, Wells D, Reid D, Sidani S. An abilities assessment instrument for elderly persons with cognitive impairment: psychometric properties and clinical utility. Journal of Nursing Measurement 1998; 6: 35-54

28. Volicer L, Hurley A, Lathi D, Kowall N. Measurement of severity in advanced Alzheimer's disease. Journal of Gerontology 1994; 49: M223-6

29. Bellelli G, Frisoni G, Bianchetti A, Trabucchi M. The Bedford Alzheimer Nursing Severity Scale for the Severely Demented: Validation Study. Alzheimer Disease and Associated Disorders 1997; 11(2): 71-7

30. Lang MM. Screening for cognitive impairment in the older adult. The Nurse Practitioner 2001; 26(11): 26, 32-4, 36-7,41