Journal of Evaluation in Clinical Practice
International Journal of Public Health Policy and Health Services Research

Journal of Evaluation in Clinical Practice ISSN 1365-2753

Methods to assess the reliability of the interRAI Acute Care: a framework to guide clinimetric testing. Part II

Nathalie I. H. Wellens MSc SLP,[1] Koen Milisen RN PhD,[3] Johan Flamaing MD PhD[4] and Philip Moons RN PhD[2]

[1]Research Associate, [2]Professor in Nursing Science, Centre for Health Services and Nursing Research, Katholieke Universiteit Leuven, Leuven, Belgium
[3]Professor in Nursing Science, Centre for Health Services and Nursing Research, Katholieke Universiteit Leuven, and Clinical Nurse Specialist, Division of Geriatric Medicine, Department of Internal Medicine, Leuven University Hospital, Leuven, Belgium
[4]Professor in Geriatric Medicine, Division of Geriatric Medicine, Department of Internal Medicine, Leuven University Hospital, Leuven, Belgium

Keywords
aged, clinimetric testing, geriatric assessment, interRAI Acute Care, reliability

Correspondence
Ms Nathalie Wellens
Centre for Health Services and Nursing Research
Katholieke Universiteit Leuven
Kapucijnenvoer 35 PB 7001
3000 Leuven
Belgium
E-mail: nathalie.wellens@med.kuleuven.be

Accepted for publication: 14 March 2011

doi:10.1111/j.1365-2753.2011.01685.x

Abstract

The interRAI Acute Care is a comprehensive geriatric assessment tool that provides a holistic picture of complex and frail hospitalized older persons. It is designed to support holistic care planning and to transfer patient data across settings. Its usefulness in clinical decision making depends on the extent to which clinicians can rely on the patient data as accurate and meaningful indicators of patients functioning. But its multidimensional character implies challenges for clinimetric testing as some of the traditional analyses techniques cannot be unconditionally applied. The objective was to present an overview of methods to examine the reliability of the interRAI Acute Care. For each line of evidence, examples of hypotheses and research questions are listed.

Introduction

Clinimetric testing of comprehensive geriatric assessment (CGA) instruments involves challenges. CGA evaluates patients' functioning on various domains. In the recent generations' CGA, multiple domains are bundled within one instrument [1,2]. This implies that clinimetric testing becomes more complex compared to testing one dimensional instruments (e.g. split-half methods are not appropriate because of the multidimensional character). The interRAI Acute Care (interRAI AC) instrument evaluates geriatric patients admitted to hospital [3]. To get a holistic picture, the patient is evaluated by 12 geriatric domains.

In a previous framework, we addressed the validity of the interRAI AC instrument [4]. Reliability is another aspect that needs scrutiny. The aim of this article is to provide a framework for testing the reliability of the interRAI AC instrument. This framework can guide future clinimetric studies on this instrument.

Methods

Instrument

The interRAI AC is tailored for older persons admitted to hospital on geriatric or non-geriatric wards [5]. Patient's history, cognition, communication, mood and behaviour, activities of daily living and instrumental activities of daily living, continence, nutrition, skin condition, medical diagnoses, health condition, medications and discharge potential are assessed over 98 items. The interRAI AC distinguishes itself not only by its standardized and extensive character but also by its structure. Four assessment periods (pre-morbid, admission, reassessment, discharge) allude to map the fluctuations of the patient's functioning over the hospital stay [5]. The pre-morbid assessment links pre-morbid and actual data and serves as a reference for the patient's capacity in the rehabilitation process. This is a unique feature which is not found in other CGA methods [1].

Once the assessment is completed, no overall score is calculated. Instead, for each domain clinical outcome measures are generated in the form of clinical assessment protocols, and scales designed to support clinical decision making for frail older patients within the acute care setting. Scales reflect severity of illness or disability in selected domains evaluating patient's current status and changes over time. Clinical assessment protocols identify areas for possible prevention or intervention of geriatric syndromes that are frequently encountered in acute care settings.

Evidence based on reliability

Reliability refers to error, both random and systematic, inherent in any measure [6,7]. It refers to the ability of an instrument to

© 2011 Blackwell Publishing Ltd,

consistently or dependably measure a target attribute. In other words, a reliable tool is one that tends to measure the same results each time it is performed and by whoever performs it [8]. In clinical use of the interRAI AC, there will always be some degree of measurement error in the administration of the item scores. Reliability testing of the interRAI AC assesses the extent to which a score is free of measurement error [9]. Reliability can be defined on a conceptual level as the degree to which 'measurements of individuals on different occasions, or by different observers, or by similar or parallel tests, produce the same or similar results' [10]. Therefore, the three key aspects of reliability are equivalence, stability and internal consistency. The reliability activities mentioned below are formulated as research questions (Q) and hypotheses (H) (Table 1).

Reliability evidence based on equivalence

One basic premise of reliability is the equivalence between assessors, that is, the extent to which various assessors obtain similar results for a specific patient's situation under identical conditions. This type of reliability is the so-called inter-rater reliability, which is important in the interest of generalizability. It allows the clinician to assume that the scores obtained by oneself or transferred by other caregivers are likely to be representative of the patient's true score. Thus, assessment results can be interpreted and applied with greater confidence [11].

To estimate inter-rater reliability, two or more trained observers examine an event simultaneously and independently record data according to an instrument's instructions [9]. Thus, to evaluate the inter-rater reliability of the interRAI AC, two independent assessors need to score the items based on the same patient's information (H_1). Even with the detailed interRAI AC manual explaining operational definitions, scoring guidelines and clinical examples, with extensive training and with equal skills, different assessors are not always in agreement about the scoring of the items being assessed [11]. Inter-rater reliability is preferably evaluated when all assessors are able to assess the patient during a single trial. Simultaneous scoring eliminates true differences in scores as a source of instrument error or carryover effect when comparing assessors' scores [11]. It can be obtained through different methods. In one method, two independent assessors accompany caregivers during the whole time interval of the assessment and score the interRAI AC items based on what they observe during the care process. A more feasible method is conducting a semi-structured interview with a patient while two independent assessors score the interRAI AC items. The assessor's function as either interviewer or observer can alter or can be determined at random each time a new patient is included. Another approach is the observation of videos. Video recordings of the patient's functioning allow multiple assessors to observe exactly the same performance [11]. In any approach, assessors must be blinded to others' results and are not allowed to discuss or to exchange information.

For the interRAI AC, analyses should be performed on item level. *The proportion observed agreement* is the ratio of exact agreement between the raters in function of the total number of assessments of the sample. It should always be reported in tandem with *Cohen's kappa coefficients* [12], a measure of agreement between assessors, corrected for chance. Unweighted Cohen's kappa is used for nominal items; weighted quadratic Cohen's kappa for ordinal items. Kappa is a score with a minimum of −1 and a maximum of 1, reflecting the agreement between raters beyond chance. The strength of agreement for the kappa coefficients is, according to Landis and Koch [13], considered as poor for kappa values below 0.40, moderate from 0.41 to 0.60, substantial from 0.61 to 0.80 and above 0.81 almost perfect. To reflect sampling error, 95% confidence intervals should be calculated. Its lower limits have to be evaluated against a clinically acceptable magnitude of 0.40 [14]. Despite the directional hypothesis that the kappa will be equal or greater than 0.40, a two-tailed test is preferable [14].

For binary items, paradoxes in agreement (e.g. a high percentage agreement versus a poor kappa agreement for the same item) can be because of bias and prevalence effects [15–17]. The prevalence of item scores affects the stability of kappa. If the ratings of the sample of patients lack variability (i.e. are homogeneous), it is unlikely that kappa will be close to the maximum of score 1. This phenomenon is independent of the sample size. Therefore, for binary items, *the prevalence index* should be calculated: the absolute value of the difference between the numbers of cases rated as positive by both raters, and the number of cases rated as negative by both raters, divided by the total number of assessments [14].

The bias index quantifies the extent to which raters disagree on the proportion of cases in a specific category of the item responses. It may be identified as the tendency of raters to have systematically different scoring patterns. Bias affects the interpretation of the magnitude of kappa. When the bias index is large, kappa tends to be higher compared to a low or absent bias [15,18]. The bias index is the absolute value of the differences between the number of cases rated as positive by rater A and negative by rater B, and the number of cases rated as negative by rater A and positive by rater B, divided by the total number of assessments. So, both indexes should be considered in interpreting the kappa values of binary items.

The complex nature of the interRAI AC instruments implies careful interpretation of the results. Within the same multidimensional data set, items are measured on nominal and ordinal level with differing scoring categories [19]. Whereas the quadratically weighted kappa coefficients tend to increase [20], the unweighted kappa coefficients decrease with the number of categories [19]. This implicates that calculating overall means of kappa coefficients of both ordinal and nominal items, or overall means of kappa coefficients of ordinal items with different number of categories should be avoided.

Some additional aspects should be borne in mind. First, in literature the *intra-class correlation coefficients* are suggested to evaluate the inter-rater agreement on ordinal items as an alternative analyses for the weighted kappa coefficients [21–23]. Second, the *Pearson product–momentum correlation* is a measure of association and not of agreement; nevertheless, it is frequently used as a reliability measure. It reflects the extent to which two observations on a group of patients can be described by a straight line. It is debated in the literature as an inappropriate measure to assess agreement because the Pearson coefficient will usually be higher than the true reliability and thus can exaggerate the impression of reliability [7,8,24]. In the extreme, two raters can correlate perfectly and yet show complete disagreement [25]. Third, the Bland–Altman method [25,26] is advocated as a better method to evaluate the agreement between two raters or between repeated measurements (cf. infra test–retest reliability). In case of the

© 2011 Blackwell Publishing Ltd

Table 1 Examples of research questions (Q) and hypotheses (H) to provide evidence on the reliability of the interRAI Acute Care

Lines of reliability
EVIDENCE BASED ON EQUIVALENCE
H_1 For each item, the score of two independent assessors is comparable.
H_2 For each pre-morbid item, the score of the interRAI AC based on patient's (or proxy's) (self-)report and the score of the interRAI HC/interRAI LTCF based on the assessment in the home care/residential setting is comparable.
H_3 For each discharge item, the score of the interRAI AC based on discharge assessment during hospitalization and the score of the interRAI HC/interRAI LTCF based on the assessment directly after hospital discharge in the home care/residential setting is comparable.
EVIDENCE BASED ON STABILITY
H_4 For all pre-morbid items, the scores obtained on repeated administrations are stable.
H_5 For admission date, the date obtained on repeated administrations is stable.
H_6 For place of residence, the scores obtained on repeated administrations are stable.
H_7 For living arrangement, the scores obtained on repeated administrations are stable.
H_8 For time since last hospital stay, the scores obtained on repeated administrations are stable.
H_9 For time spend in emergency room, the scores obtained on repeated administrations are stable.
H_{10} For surgery, the scores obtained on repeated administrations are stable.
H_{11} For admission (working) diagnoses, the diagnoses obtained on repeated administrations are stable.
H_{12} For discharge (working) diagnoses, the diagnoses obtained on repeated administrations are stable.
H_{13} For admission medications, the list obtained on repeated administrations is stable.
H_{14} For discharge medications, the list obtained on repeated administrations is stable.
H_{15} For advanced directives, the scores obtained on repeated administrations are stable.
H_{16} For community services prior to admission, the list obtained on repeated administrations is stable.
H_{17} For date of discharge, the date obtained on repeated administrations is stable.
H_{18} For discharge destination, the scores obtained on repeated administrations are stable.
H_{19} For hearing, the scores obtained on repeated administrations are stable.
H_{20} For vision, the scores obtained on repeated administrations are stable.
H_{21} For weight and height, the scores obtained on repeated administrations are stable.
H_{22} For making oneself understood, the scores obtained on repeated administrations are stable.
H_{23} For ability to understand others, the scores obtained on repeated administrations are stable.
H_{24} For required treatments and procedures after discharge, the scores obtained on repeated administrations are stable.
H_{25} For each item, the scores obtained on repeated administrations of one assessor are comparable.
EVIDENCE BASED ON INTERNAL CONSISTENCY
H_{26} The Depression Rating Scale is internal consistent.
H_{27} The Communication Scale is internal consistent.
H_{28} Each item of the Depression Rating Scale is correlated with the total score of the Depression Rating Scale.
H_{29} Each item of the Communication Scale is correlated with the total score of the Communication Scale.
H_{30} Each item of the Cognitive Performance Scale is correlated with the total score of the Cognitive Performance Scale.
H_{31} Each item of the Activities of Daily Living Scale is correlated with the total score of the Activities of Daily Living Scale.
H_{32} Each item of the Instrumental Activities of Daily Living Scale is correlated with the total score of the Instrumental Activities of Daily Living Scale.
H_{33} Each item of the Pain Scale is correlated with the total score of the Pain Scale.
H_{34} Deleting one of the items of the Depression Rating Scale will not greatly increase the overall reliability of the Depression Rating Scale.
H_{35} Deleting one of the items of the Communication Scale will not greatly increase the overall reliability of the Communication Scale.
H_{36} Each item of the Depression Rating Scale correlates with the total score of the Depression Rating Scale and does not correlate with the total score of any other scale.
H_{37} Each item of the Communication Scale correlates with the total score of the Communication Scale and does not correlate with the total score of any other scale.
H_{38} Each item of the Cognitive Performance Scale correlates with the total score of the Cognitive Performance Scale and does not correlate with the total score of any other scale.
H_{39} Each item of the Activities of Daily Living Scale correlates with the total score of the Activities of Daily Living Scale and does not correlate with the total score of any other scale.
H_{40} Each item of the Instrumental Activities of Daily Living Scale correlates with the total score of the Instrumental Activities of Daily Living Scale and does not correlate with the total score of any other scale.
H_{41} Each item of the Pain Scale correlates with the total score of the Pain Scale and does not correlate with the total score of any other scale.

interRAI HC, interRAI Home Care; interRAI LTCF, interRAI Long Term Care Facility.

© 2011 Blackwell Publishing Ltd

interRAI AC, this method could not be used because no total score is calculated.

In addition to the traditional inter-rater reliability, for the interRAI AC, two alternative procedures can be tested. First, next to inter-rater reliability on item level, agreement in diagnostics or in clinical decision making based on the interRAI AC findings (e.g. clinical assessment protocols) could be examined. Second, because the interRAI portfolio aims the transfer of accurate and reliable data across care settings, the agreement between assessors in different settings can be evaluated. For the pre-morbid assessment of the interRAI AC, agreement between the pre-morbid data based on the report of the patient or proxy at hospital admission and the pre-morbid functioning of the patient as scored by the caregiver in the home care (interRAI Home Care, interRAI HC) or residential care (interRAI Long Term Care Facility, interRAI LTCF) can be evaluated (H_2). Similarly, for the discharge assessment, agreement between the scores of the discharge data of the interRAI AC and the scores of the interRAI Home Care or the interRAI Long Term Care Facility as completed directly after hospital discharge can be evaluated (H_3). Reliability of transferred data is an important feature in the interest of continuity of care.

Reliability evidence based on stability

One basic premise of reliability is the stability of the assessment instrument, that is, the extent to which similar results are obtained on repeated administrations [11]. There are two types of reliability evidence based on stability: test–retest reliability and intra-rater reliability.

Test–retest reliability is used to evaluate if an instrument is capable of measuring patient data with consistency [11]. Test–retest procedures involve administering the instrument under study to the same group at separate occasions – keeping all assessment conditions as constant as possible – and comparing the item scores [9]. The agreement between the two scores reflects the stability of the instrument [27]. Test–retest reliability is mostly used for self-rated tests [10]. In the case of the interRAI AC, patients (or proxies) are asked to report the pre-morbid functioning. Hence, test–retest reliability can be tested for all *pre-morbid items* (H_4). Furthermore, test–retest reliability is only relevant for attributes that are not expected to change over time [11,27]. Hence, this aspect of reliability is hardly measurable in an acute setting because of the variable character of acute symptoms, especially in older populations [28,29]. Along these lines, test–retest reliability of *the clinical items during hospital stay* is not relevant for the interRAI AC, because its fundamental goal is to register small changes in the patient's functioning and needs during hospitalization. Only for administrative items (e.g. living arrangement, time since last hospital stay; H_5–H_{18}), and for some clinical interRAI AC items for which we can assume that these are unlikely to fluctuate much from day to day (e.g. hearing and vision; H_{19}–H_{24}), stability measurements are feasible. Assessment within the context of an acute setting implies some additional remarks. Because of the short length of stay of hospitalized frail older persons, the time interval between the two registrations needs to be relatively short [10]. This may impact the strength of agreement because the assessor may remember the earlier administered scores and/or patients may remember their responses (e.g. on pre-morbid items). On the other hand, differences in agreement may have occurred because the patient's responses on self-report items (e.g. pre-morbid items) may be inconsistent [30,31].

An assessor may apply slightly different standards or rating patterns from day to day. Therefore, **the intra-rater reliability** measures variation that occurs within an assessor as a result of multiple exposures to exactly the same patient's situation (H_{25}) [10]. This could be tested experimentally by videotaping fragments of patients functioning and a semi-structured interview and having the assessor do two interRAI AC ratings based on the tapes with a time interval of at least 1 week or two apart to minimize rater bias (e.g. memory effects) [11].

No assessor yields exactly the same result from assessment to assessment. Therefore, it is necessary to determine a standard for acceptable level of error [8]. Similar reliability coefficients and levels of acceptability can be used as described above for the inter-rater reliability [9].

Whereas test–retest reliability evaluates mainly the stability of the assessment instrument, intra-rater reliability evaluates mainly the stability of the rater. Both use a design with repeated measures. For instruments like the interRAI AC, where the skills of the assessor are relevant to the accuracy of the assessment scoring, both test–retest reliability and intra-rater reliability are essentially the same estimate. Thus, for this instrument, the stability of the rater and the assessment cannot be separated out [11]. Moreover, in case that high inter-rater reliability (cf. supra reliability evidence based on equivalence) is demonstrated, tests of intra-rater reliability may be unnecessary [10].

Reliability evidence based on internal consistency

The premise for testing internal consistency is that groups of items are thought to measure different aspects of the same concept. An instrument is internally consistent or homogeneous if its items or subparts measure (various aspects of) the same dimension or fit together conceptually [9]. The usual way to look at this type of reliability is based on the idea that individual items (or set of items) should produce results consistent with the overall questionnaire [32]. Testing of the internal consistency is a popular and widely used technique of testing the reliability, because it can be computed from routine administration, without the requirement of two or more administrations [33]. Item homogeneity can be quantified in various ways including the split-half technique, Cronbach's alpha, the Kuder–Richardson formula 20 and corrected item-total correlation.

The split-half technique randomly divides the items into two subscales, which are then correlated [10]. Because of the multidimensionality of the interRAI AC, this measure is not appropriate.

Cronbach's alpha is a method that splits the interRAI AC data into two sets, in every possible way, and computes the average of all correlation coefficient for each split [10]. In the strict interpretation of both definitions, this type of reliability is not appropriate to the interRAI AC for three reasons [33,34]. First, because of its multidimensional character, the different sub-domains of the CGA instrument are – per definition – not homogeneous. Second, the interRAI AC items are not intended to be summed to result in an overall score. Third, Cronbach's alpha is strongly affected by the number of instrument items [6,10,35]. Therefore, the value of alpha is a priori hypothesized to be relatively high, because the

© 2011 Blackwell Publishing Ltd

interRAI AC consist of 98 items. A high alpha value will simply reflect the length of the interRAI AC, not its composition.

Alternatively, 'a second interpretation of alpha is that it measures unidimensionality, or the extent to which the scale measures one underlying factor or construct (. . .) Cronbach suggested (1951) that if several factors exist then the formula should be applied separately to items relating to different factors' (p. 675) [32]. In other words, the internal consistency could be measured for the embedded scales of the interRAI AC, separately. Cronbach's alpha may only be used in cases of summated scales, because each item of the scale must be rated similarly and must contribute equally to the total score [36]. Some of the embedded scales of the interRAI AC are summated scales; others are based on algorithms that combine items. Concretely, the internal consistency could only be estimated for the Depression Rating Scale (DRS, a summated scale providing information on signs of possible depression, based on three ordinal items of three scoring categories) and the Communication Scale (CS, a summated scale indicating the level of communication skills based on two ordinal items of five scoring categories; H_{26}, H_{27}). All the other embedded scales (i.e. Cognitive Performance Scale, Activities of Daily Living Hierarchy, Pain Scale, Instrumental Activities of Daily Living Scale) are calculated by algorithms and/or based on a summation of items with differing scoring categories. Thus, Cronbach's alpha is applicable for the DRS and CS, but is inappropriate for the other scales. A value of 0.8 is in general acceptable for Cronbach's alpha. Lower values would indicate an unreliable scale. These standards are differentiated by Kline noting that the cut-off of 0.8 is appropriate for cognitive tests (e.g. DRS), and 0.7 is more suitable for ability tests (e.g. CS) [37]. Additional analyses include *Corrected Item-Total Correlation* and *Cronbach's alpha if item deleted*. The former reflects correlations between each item and the total score of the scale being tested (H_{28}–H_{33}). Items with values below 0.3 are potentially problematic in the light of internal consistency. The latter will tell if removing one of the scale items will improve the overall reliability of the embedded scale being tested. Values greater than the overall reliability indicate that if a particular item is not included in the calculation, the overall reliability of the scale will improve (i.e. the alpha value will increase; H_{34}, H_{35}). As mentioned earlier, calculations based on Cronbach's alpha can be applied to summated scales only.

The Kuder–Richardson 20 is appropriate for scales with binary items [6]. Because the interRAI AC has items on differing measurement levels, this technique is not applicable.

For multidimensional tools which comprise a number of subscales more sophisticated analytical techniques are also suggested. In these techniques, an item of a subscale is correlated with its subscale total, and with the totals of all other subscales. It is assumed that the item correlates well with the subscale that it belongs to, and not with any other subscale (H_{36}–H_{41}). If the correlation of the item is higher with another subscale or if it correlates on two or more subscales, then it is likely that it may be tapping another dimension than intended in the developmental stage of the tool. Hence, the algorithm should be rewritten or eliminated.

Discussion

The interRAI AC is designed to support holistic care planning and to transfer patient data across settings, based on standardized CGA [1,38]. The usefulness of the interRAI AC assessment in clinical decision making depends on the extent to which clinicians can rely on the patient data as accurate and meaningful indicators of patients functioning [11]. There is also a growing need for uniformity in assessment for research [39], monitoring and regulatory purposes. On all these levels, reliability testing is important, regardless of the end user being a clinician, a researcher, a hospital manager or a policy maker. The aim of the current paper was to supply a structured framework of the clinimetric tests to evaluate different lines of reliability of the interRAI AC.

Reliability is not a fixed property of an instrument; it cannot be conceived as a property that the interRAI AC possess or does not possess, rather it will have a certain degree of reliability when applied to certain populations under certain conditions [10,40]. In this regard, future clinimetric studies should extensively and transparently report all testing conditions: the type of reliability coefficient (e.g. intra-class correlation coefficient, kappa, percentage observed agreement), the sample size, the characteristics of the group being assessed (e.g. age, gender, ethnicity, cognitive functioning, inclusion and exclusion criteria), the characteristics of the acute setting (e.g. length of stay, ward type), the characteristics of the assessors (e.g. type of health profession, years of experience), the RAI training (e.g. duration of the training, type of training), the consulted sources of information (e.g. daily clinical observation, semi-structured patient interview, family, medical file) and the circumstances under which the assessment is carried out should be described in detail. Accordingly, we cannot automatically assume that estimates from one study can be generalized to other assessors, clinical contexts or types of patients [11]. Preferably, future reports on reliability studies should address the proposed guidelines for reporting reliability and agreement studies [41].

Prior research on the reliability of the interRAI AC was restricted to the inter-rater reliability [3,5], and was limited to the earlier draft versions of the instrument. Future research should evaluate multiple dimensions of reliability, should use a combination of techniques of analyses [41] and should test the latest version of the interRAI AC. Until now, the interRAI AC is only in routine clinical use in Australia. In Belgium, a governmental project examines and supports a nationwide implementation process. To the best of our knowledge, start-ups and demonstrations are occurring or planned in Canada, Finland, Iceland, Italy, Norway, Singapore and Spain. Because the contexts of acute care can vary across nations, hospital types, local legislation, etc., ideally reliability testing in tandem with validity [4] testing should be carried out in these different contexts in order to obtain a wide body of evidence. The actual testing of research questions and hypotheses is currently in progress [42] and will be continued in future research.

References

1. Wellens, N. I. H., Deschodt, M., Flamaing, J., Moons, P., Boonen, S., Boman, X., Gosset, C., Petermans, J. & Milisen, K. (2011) First-generation versus third-generation comprehensive geriatric assessment instruments in the acute hospital setting: a comparison of the minimum geriatric screening tools (MGST) and the interRAI Acute Care (interRAI AC). *The Journal of Nutrition, Health & Aging* (in press).
2. Bernabei, R., Landi, F., Onder, G., Liperoti, R. & Garnbassi, G. (2008) Second and third generation assessment instruments: the birth of standardization in geriatric care. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 63 (3), 308–313.

 © 2011 Blackwell Publishing Ltd

3. Gray, L. C., Bernabei, R., Berg, K., Finne-Soveri, H., Fries, B., Hirdes, J. P., Jónsson, P., Morris, J. N., Steel, K. & Arino-Blasco, S. (2008) Standardizing assessment of elderly people in acute care: the interRAI Acute Care instrument. *Journal of the American Geriatrics Society*, 56 (3), 536–541.
4. Wellens, N. I. H., Milisen, K., Flamaing, J. & Moons, P. (2010) Methods to assess the validity of the interRAI Acute Care: a framework to guide clinimetric testing. *Journal of Evaluation in Clinical Practice* (Epub ahead of print). doi: 10.1111/j.1365-2753.2010.01571.x.
5. Carpenter, G. I., Teare, G. F., Steel, K., Berg, K., Murphy, K., Bjornson, J., Jonsson, P. V. & Hirdes, J. P. (2001) A new assessment for elders admitted to acute care: reliability of the MDS-AC. *Aging (Milano)*, 13 (4), 316–330.
6. Nunnally, J. C. (1978) Psychometric Theory, 2nd edn. New York: McGraw-Hill.
7. Streiner, D. L. & Norman, G. R. (2006) 'Precision' and 'accuracy': two terms that are neither. *Journal of Clinical Epidemiology*, 59 (4), 327–330.
8. Bannigan, K. & Watson, R. (2009) Reliability and validity in a nutshell. *Journal of Clinical Nursing*, 18 (23), 3237–3243.
9. Polit, D. F. & Beck, C. T. (2008) Nursing Research: Generating and Assessing Evidence for Nursing Practice, 8th edn. Philadelphia, PA: Lippincott Williams & Wilkins.
10. Streiner, D. L. & Norman, G. R. (2008) Health Measurement Scales: A Practical Guide to Their Development and Use, 4th edn. Oxford: Oxford University Press.
11. Portney, L. G. & Watkins, M. P. (2009) Foundations of Clinical Research Applications to Practice, 3rd edn. Upper Saddle River, NJ: Pearson/Prentice Hall.
12. Cicchetti, D. V. & Sparrow, S. A. (1981) Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. *American Journal of Mental Deficiency*, 86 (2), 127–137.
13. Landis, J. R. & Koch, G. G. (1977) The measurement of observer agreement for categorical data. *Biometrics*, 33 (1), 159–174.
14. Sim, J. & Wright, C. C. (2005) The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical Therapy*, 85 (3), 257–268.
15. Byrt, T., Bishop, J. & Carlin, J. B. (1993) Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, 46 (5), 423–429.
16. Hoehler, F. K. (2000) Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. *Journal of Clinical Epidemiology*, 53 (3), 499–503.
17. Vach, W. (2005) The dependence of Cohen's kappa on the prevalence does not matter. *Journal of Clinical Epidemiology*, 58 (7), 655–661.
18. 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.
19. Brenner, H. & Kliebsch, U. (1996) Dependence of weighted kappa coefficients on the number of categories. *Epidemiology*, 7 (2), 199–202.
20. Maclure, M. & Willett, W. C. (1987) Misinterpretation and misuse of the kappa statistic. *American Journal of Epidemiology*, 126 (2), 161–169.
21. Field, A. (2005) Intraclass correlation. In Encyclopedia of Behavioral Studies (eds B. Everitt & D. C. Howell), pp. 948–954. New York: Wiley.
22. Fleiss, J. L. & Cohen, J. (1973) Equivalence of weighted kappa and intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33 (3), 613–619.
23. Laschinger, H. K. (1992) Intraclass correlations as estimates of interrater reliability in nursing research. *Western Journal of Nursing Research*, 14 (2), 246–251.
24. Griffiths, P. & Murrells, T. (2010) Reliability assessment and approaches to determining agreement between measurements: classic methods paper. *International Journal of Nursing Studies*, 47 (8), 937–938.
25. Bland, J. M. & Altman, D. G. (1986) Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, 1 (8476), 307–310.
26. Bland, J. M. & Altman, D. G. (2003) Applying the right statistics: analyses of measurement studies. *Ultrasound in Obstetrics & Gynecology*, 22 (1), 85–93.
27. Devon, H. A., Block, M. E., Moyle-Wright, P., Ernst, D. M., Hayden, S. J., Lazzara, D. J., Savoy, S. M. & Kostas-Polston, E. (2007) A psychometric toolbox for testing validity and reliability. *Journal of Nursing Scholarship*, 39, 155–164.
28. Hirsch, C. H., Sommers, L., Olsen, A., Mullen, L. & Winograd, C. H. (1990) The natural history of functional morbidity in hospitalized older patients. *Journal of the American Geriatrics Society*, 38 (12), 1296–1303.
29. Hoogerduijn, J. G., Schuurmans, M. J., Duijnstee, M. S., de Rooij, S. E. & Grypdonck, M. F. (2007) A systematic review of predictors and screening instruments to identify older hospitalized patients at risk for functional decline. *Journal of Clinical Nursing*, 16 (1), 46–57.
30. Colsher, P. L. & Wallace, R. B. (1989) Data quality and age: health and psychobehavioral correlates of item nonresponse and inconsistent responses. *Journal of Gerontology*, 44 (2), 45–52.
31. Sherbourne, C. D. & Meredith, L. S. (1992) Quality of self-report data: a comparison of older and younger chronically ill patients. *Journal of Gerontology*, 47 (4), S204–S211.
32. Field, A. P. (2009) Discovering Statistics Using SPSS (and Sex and Drugs and Rock 'n' Roll), 3rd edn. Los Angeles, CA: Sage.
33. Streiner, D. L. (2003) Starting at the beginning: an introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, 80 (1), 99–103.
34. Streiner, D. L. (2003) Being inconsistent about consistency: when coefficient alpha does and doesn't matter. *Journal of Personality Assessment*, 80 (3), 217–222.
35. Cortina, J. M. (1993) What Is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78, 98–104.
36. Carmines, E. G. (1979) Reliability and Validity Assessment. Beverly Hills, CA: Sage Publications.
37. Kline, P. (1999) The Handbook of Psychological Testing. London: Routledge.
38. Gray, L. C., Berg, K., Fries, B. E., Henrard, J. C., Hirdes, J. P., Steel, K. & Morris, J. N. (2009) Sharing clinical information across care settings: the birth of an integrated assessment system. *BMC Health Services Research*, 29 (9), 71.
39. Van Craen, K., Braes, T., Wellens, N., Denhaerynck , K., Flamaing, J., Moons, P., Boonen, S., Gosset, C., Petermans, J. & Milisen, K. (2010) The effectiveness of inpatient geriatric evaluation and management units: a systematic review and meta-analysis. *Journal of the American Geriatrics Society*, 58 (1), 83–92.
40. Wilkinson, L. (1999) Statistical methods in psychology journals – guidelines and explanations. *American Psychologist*, 54 (8), 594–604.
41. Kottner, J., Audige, L., Brorson, S., Donner, A., Gajewski, B. J., Hrobjartsson, A., Roberts, C., Shoukri, M. & Streiner, D. L. (2011) Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *Journal of Clinical Epidemiology*, 64 (1), 96–106.
42. Wellens, N. I. H., Deschodt, M., Boonen, S., Flamaing, J., Gray, L., Moons, P. & Milisen, K. (2011) Validity of the interRAI Acute Care based on test content: a multi-center study. *Aging Clinical and Experimental Research* (in press).

© 2011 Blackwell Publishing Ltd