Unfolding the phenomenon of interrater agreement: a multicomponent approach for in-depth examination was proposed.

Slaug, Björn; Schilling, Oliver; Helle, Tina; Iwarsson, Susanne; Carlsson, Gunilla; Brandt, Åse

Published in:
Journal of Clinical Epidemiology

DOI:
10.1016/j.jclinepi.2012.02.016

2012

Unfolding the phenomenon of inter-rater agreement: a multi-component approach for in-depth examination was proposed

Björn Slaug, Oliver Schilling, Tina Helle, Susanne Iwarsson, Gunilla Carlsson and Åse Brandt

Full name, degree, title and affiliation:
Slaug, Björn, BA[1]
Schilling, Oliver, PhD[2]
Helle, Tina, MSc, Reg. OT[1,4]
Iwarsson, Susanne, PhD, Professor, Reg. OT[1]
Carlsson, Gunilla, PhD, Reg. OT[1]
Brandt, Åse, PhD, MPH, Reg. OT[3]
[1)] *Department of Health Sciences, Faculty of Medicine, Lund University, Sweden*
[2)] *Department of Psychological Ageing Research, Institute of Psychology, University of Heidelberg, Germany*
[3)],*The National Board of Social Services, Odense, Denmark Århus, Denmark*
[4)] *Department of Occupational Therapy, University College Northern Jutland, Ålborg, Denmark*

Corresponding author:
Björn Slaug
Department of Health Sciences, Faculty of Medicine, Lund University
Box 157
SE-221 00 Lund
SWEDEN
Tel: +46-46-2221838, Fax: +46-46-2221959
E-mail: bjorn.slaug@med.lu.se

ABSTRACT

Objective: The overall objective was to unfold the phenomenon of inter-rater agreement: to identify potential sources of variation in agreement data and to explore how they can be statistically accounted for. The ultimate aim was to propose recommendations for in-depth examination of agreement, in order to improve the reliability of assessment instruments.

Study Design and Setting: Utilizing a sample where 10 rater pairs had assessed the presence/absence of 188 environmental barriers by a systematic rating form, a raters × items dataset was generated (N=1,880). In addition to common agreement indices, relative shares of agreement variation were calculated. Multilevel regression analysis was carried out, using rater and item characteristics as predictors of agreement variation.

Results: Following a conceptual decomposition, the agreement variation was statistically disentangled into relative shares. The raters accounted for 6-11%, the items for 32-33% and the residual for 57-60% of the variation. Multilevel regression analysis showed barrier prevalence and raters' familiarity with using standardized instruments to have the strongest impact on agreement.

Conclusion: Supported by a conceptual analysis we propose an approach of in-depth examination of agreement variation, as a strategy for increasing the level of inter-rater agreement. By identifying and limiting the most important sources of disagreement, ultimately instrument reliability can be improved.

Keywords: inter-rater, reliability, agreement, kappa, methodology, recommendations

AUTHORS' NOTE

This study was accomplished within the context of the Centre for Ageing and Supportive Environments (CASE), at Lund University, by funding from the Ribbing Foundation, Lund Sweden and by the Swedish Council for Working Life and Social Research.

INTRODUCTION

In research and practice contexts aiming to develop supportive environments for health, complex assessments targeting personal as well as environmental components are necessary. Such assessments involving raters imply a diversity of challenges, not the least as concerns *instrument reliability*. With instrument reliability we refer in a broad sense to features of stability and consistency in instrument use by different raters [1]. Measuring the level of inter-rater agreement informs about the extent to which different raters essentially make the same assessments [2-4]. This fundamental aspect of instrument reliability concerns the stability of assessments across different raters. For an instrument to be regarded as reliable though, it is equally important that there is a consistency among raters with respect to assessment variability [2-4]. To achieve high reliability, in addition to high agreement it must also be possible to detect true variation by means of the assessments [5].

Typically, on the path towards reliable instrument use researchers accomplish inter-rater agreement/reliability studies. However, as shown by such studies within health sciences [6-8] the link between statistical analysis and conclusions regarding instrument reliability is multi faceted and often difficult to interpret. Consequently, there is a need for more in-depth examination than conventionally applied. As recently pointed out [9], one reason for scarce knowledge about the reliability of assessment instruments is the lack of widely accepted standards for reporting agreement/reliability studies. To further contribute to the set of recommendations proposed in Kottner et al.'s study, we will focus on advancing the methods for examining data generated in agreement/reliability studies. The most valuable and useful information, we will argue, may come from disentangling and analysing variation in agreement data. In order to pursue this, we will explore the phenomenon of agreement conceptually, thereby relating it to how variation in agreement can be statistically accounted

for. To demonstrate the benefit of such examination, we will utilize data collected in realistic rating situations in the homes of persons with functional limitations, from a previous agreement/reliability study concerning housing accessibility assessments [6].

One major challenge in striving for reliability concerns the complexity of the instrument used. With a composite instrument, there may be items differing in the degree they discriminate between phenomena. For example, if an item represents presence or absence of a characteristic that in reality rarely occurs at all, the discriminating power of such an item is minor. That is, it is easy to agree on absence if the discriminating ability will seldom be put to the test. In addition, there may be a mix of items in terms of administration differences [10]. Some items may only require "straight and simple" measuring with a measurement device, while others may require observation of perceptual phenomena, and yet others may rely on evaluative judgement, largely depending on the competence and experience of the raters. Moreover, although it is assumed that an assessment which is defined in a certain way in essence is the same from case to case, in reality two cases are never completely identical, as the contextual situations differ.

Evidently, in any inter-rater agreement study some degree of disagreement is likely to occur. Without assumptions how to explain the occurrence of disagreement, conclusions attempting to generalize the results will be weak. Therefore it would improve the explanatory potential to carefully consider what can reasonably be assumed to be possible sources of disagreement prior to carrying out the study, and include variables that cover these sources when collecting data [11]. Examples of sources of disagreement can be rater background or contextual particularities of the rating situations such as weather or lighting conditions. Thus, the strength of conclusions is highly dependent on whether the most relevant sources of

disagreement were accounted for or not; to understand the level of agreement reached, it is crucial to be able to sort out the most likely reasons why disagreement occurs. On a theoretical level the main sources of disagreement could be described in abstract terms by a conceptual decomposition of agreement and further specified by identifying relevant aspects of the components recognised. On a concrete level, managing a whole set of agreement data, variation in agreement as measured by statistical indices could be partitioned and structured accordingly. Hence, by applying these theoretical principles to data analysis, a more systematic and in-depth examination of agreement variation can be accomplished.

Study objectives

The overall objective of this study was to unfold the phenomenon of inter-rater agreement, conceptually and statistically. The specific aims were to identify potential sources of variation influencing inter-rater agreement, and to explore how they can be statistically accounted for. Thus, we aim to present analytical approaches to decompose and explain impacts on inter-rater agreement. The ultimate aim of the study was to come up with recommendations for in-depth examination of inter-rater agreement, in order to improve the reliability of assessment instruments, particularly within health sciences.

CONCEPTUAL ANALYSIS

The mere fact that two raters *apparently* agree, in registering identical responses on a given question is not in itself sufficient to conclude that they *truly* agree. In other words, a distinction can be made between "true" agreement and "apparent" agreement. The fact that two raters come to identical or different responses though seemingly answering the same question, may be influenced by a number of factors: the clarity/obscurity of the instructions,

the conditions under which the raters make the assessments, the tools that they have at their disposal, the characteristics of the phenomena assessed, etc [9]. For example, one rater may have misunderstood the instructions, but nevertheless in some cases his responses may be the same as if he had followed the instructions correctly [12]. This is not only conceptually important, but also has the implication that to capture the full meaning of agreement, more information than just the assessment responses need to be taken into consideration.

A first intuitive definition of inter-rater agreement could be: "if both rater A and rater B say X is the case, they agree", meaning that they both come to identical assessment results. "X is the case" should be seen as a general formulation, representing a statement of the true state of matters. However, even if they both say "X is the case", but the phenomenon they are assessing and to which the "X is the case" statement refers, differ in its characteristics relevant for the assessment, it would not make sense to define it as agreement; "X is the case" would have different meanings for A and B. A first qualification of the definition would then be "If both rater A and rater B say X is the case concerning Y, they agree", meaning that they assess the same phenomenon, or phenomena that are identical in characteristics relevant for the assessment. Nevertheless, also this definition falls short after some reflection. If for instance, the assessment requires observation during certain lighting or temperature conditions and these are present only for rater A but not for rater B, then it would also be questionable to define it as agreement. So a second qualification would be "If both rater A and rater B say X is the case concerning Y under condition Z, they agree". With "condition Z" is understood all contextual circumstances of the assessment, including occasion in time (see also comparable definition of inter-rater agreement [9:Appendix]).

Conceptual diversification of agreement into three main components

Following this reasoning, and as implicated already in early studies in the field [13] and even more clearly stated recently [9], we propose to diversify the concept of agreement into three main components, corresponding to the elements of the definition suggested above.

1) The rater component, covering personal characteristics specific to the individual pair of raters, such as their professional experience and inclination to be liberal or severe in their assessment.
2) The item component, covering characteristics of the phenomenon to be rated as formulated by the item definition, which may call for different assessment methods such as ocular observation, evaluative judgement or the use of a measurement device.
3) The context component, covering characteristics and attributes of varying contextual conditions, under which the assessment task is carried out, denoting particularities of the whole assessment situation, such as the time of the day, season of the year, lighting, weather conditions, presence of other people, instructions in the instrument manual etc.

In Figure 1 we have listed different characteristics subsumed under each component that are reasonable to assume having influence on inter-rater agreement (for a similar overview, e.g. [14]). Thus *characteristics* denote concrete representations of what we in an abstract sense term the *aspects* of the components.

Raters	Items	Contexts
Relevant professional experience	*Type / mode of item rating*	General instructions of the manual
Familiarity with the use of standardized rating instruments	*Prevalence of phenomenon referred to by item*	Precision of rating tools / devices
Inclination to be liberal or severe in rating	Item rating scale	Time of the day
Extent of rating education / training	Precision / wording of item definition	Time frame of rating
Closeness in time of rating education / training	Item information amount necessary for rating	Weather conditions
Rater competence	Nature of phenomenon referred to by item definition	Lighting conditions
		Season of the year
		Temperature
		Distractions in the rating environment

Note: Characteristics used in the data analysis of the current study are marked with italic letters.

Figure 1. The three main components of agreement and examples of characteristics subsumed.

The underlying assumption of this diversification is that in reality agreement is not something absolute, which is either at hand or not, but something relative to the aspects of the components. That means, a claim of attained true agreement implies an inference of impartial influence from any of the components. However, that represents an ideal situation that may occur in theory, but in realistic rating situations apparent agreement will inevitably cover a variation in the aspects of the components. When represented by data where agreement is recorded as pair-wise assessments that are either identical or not, this dimension may seem unaccounted for. Yet, each observation of pair-wise assessments can be regarded as a sample taken from a universe of possible observations with varying aspects of the components, and a

reasonably large a set of observations can therefore be considered to be representative of this variation [15-16]. Moreover, assuming that variation in agreement can be seen as a reflection of the influence of the components, statistical techniques for analysing variance should be particularly appropriate to apply. Thus, it would be valuable to know the share of variation accounted for by each component. Such an analysis would give guidance and an indication as to the domain where the most influential sources of disagreement are likely to be found.

However, as such an analysis only point to the relative importance of each component, there is a need for further examination, which can indicate the magnitude and direction of influence of more specific aspects of the components as well. That means, each observation of pair-wise assessments has to be linked to characteristics of the raters, items and contexts, respectively, which can serve as representations of relevant aspects of the components, and thus also constitute sources of disagreement. Such analysis has the potential of providing very detailed and specific information on the most important sources of disagreement, and thus where to put the most efforts in order to improve the level of agreement, and ultimately, the reliability of assessment instruments.

STATISTICAL ANALYSIS

Data used

For this study, a dataset generated from a previous agreement/reliability study [6] was utilized. The Helle et al. study was a cross-Nordic project, heading towards reliable accessibility assessments of the physical environment. The dataset was generated from a sample where 10 rater pairs had assessed 8-14 different cases each (in total 106 cases), and included data from Denmark, Finland, Iceland and Sweden. Each case concerned a unique

dwelling and comprised pair-wise dichotomous assessments of presence/absence of 188 physical environmental barriers (henceforth only called barriers) in the home and the immediate outdoor environment, as defined by the Housing Enabler instrument [17]. To fit the design of the current study, the dataset was re-structured in a raters × items matrix. That is, for each constellation of rater pair and item the cell frequencies were computed by cross-tabulating the pair-wise assessments of presence/absence of the barriers. As a result, a dataset comprising 1,880 observations (10 rater pairs × 188 items) was generated.

A sampling strategy following joint principles was applied in all four countries. The ten rater pairs were instructed to organise and strive for the largest possible diversity concerning type of dwelling. All raters had completed a four-day course, conducted by the same course leaders and following the same format. In Sweden the raters had up to three years of experience from using the instrument and possessed previous experiences from participation in a research project. In contrast, the Finnish and Danish raters had no previous experience from using Housing Enabler, and only a few of them were used to employ standardised assessments. The data collection was performed at home visits, over a period of two months. Each case was assessed independently by each of the two raters of a rater pair within one week. For further details see [6].

In addition, Housing Enabler data (N=1,150) from a European, interdisciplinary research project, the ENABLE-AGE, were used to provide a non-sample dependent estimate of barrier prevalence, assumed to reflect a common prevalence of barriers in ordinary dwellings. Details of the project have been published elsewhere (see e.g. [18-19]).

Predictors of agreement variation

Restricted by the original data where contextual characteristics were not systematically controlled for, the current study only allowed for predictors adherent to raters and items. Characteristics attributable to raters were retained by two variables: "Professional housing adaptation experience" and "Familiarity with professional use of standardized assessment instruments". These two variables were given the value of '1' when both raters in a pair had a record of at least one year of professional experience/familiarity. Otherwise if only one or none of the raters had such professional experience/familiarity recorded, the variables were given the value '0'. Characteristics attributable to items were likewise retained in two variables: "Barrier assessment type" and "Barrier prevalence estimate". Barrier assessment type categorizes the items in three different types according to their type or mode of assessment [10] on how to discriminate between presence/absence. Those items assessed by means of a rule or tape measure were classified as *Measurable*, those assessed by means of perception-based judgments as *Obvious by observation* and those assessed by evaluation-based judgments as *Evaluabl*e. Barrier prevalence estimate was defined by utilizing the Housing Enabler data from the ENABLE-AGE project. See Table 1 for a description of the predictors in our dataset.

Table 1. Description of variables used as predictors of agreement variation.

Component of agreement Characteristic used as predictor for agreement variation	
Raters	*N=10*
Housing adaptation experience: Both raters, n (%)	8 (80.0)
Familiarity with standardized instruments: Both raters, n (%)	2 (20.0)
Items	*N=188*
Barrier assessment type: Measurable, n (%)	72 (38.3)
Barrier assessment type: Obvious by observation, n (%)	48 (25.5)
Barrier assessment type: Evaluable, n (%)	68 (36.2)
Barrier prevalence estimate[a], mean (SD)	33.2 (27.2)

[a] Barrier prevalence is estimated as the occurrence in the ENABLE-AGE sample.

Agreement indices

For each constellation of rater pair and item (N=1,880), agreement indices *Observed agreement (P_o)* and *Kappa (κ)* were calculated (for formulas, see Appendix). The P_o and κ indices were then treated as variables under examination in all the subsequent analyses. We selected these two agreement indices as they are the most common and recommended [20-22] for measuring inter-rater agreement.

Level of agreement analysis

The index means were calculated for the 1,880 observations, in total and by rater and item characteristics. Barrier prevalence estimate was classified in five distributional categories, 0-20%, 21-40%, 41-60%, 61-80% and 81-100%. Moreover, following a suggestion that a balanced prevalence near 50% [23] should be targeted for the "fairest" assessment of the level of agreement, the items with prevalence estimate in the interval 41-60% were highlighted in an extended analysis. Due to the results of the predictors of agreement variation analysis described below, splitting the mean levels by rater and item characteristics did not include housing adaptation experience.

Shares of agreement variation analysis

Based on the conceptual analysis, we defined a "*Shares of agreement variation formula"*, which intends to disentangle the contribution of the components (raters, items and contexts) into relative shares. Given our agreement indices (P_o and κ) that are systematically arranged in a raters × items matrix, the basic decomposition of the overall agreement variation would employ the variation due to both components and the residual variation. That is, the agreement values' total sum of squares (SST) could be decomposed into the sums of squares

due to rater pair variation (SSR), due to item variation (SSI), and the residual sum of squares (SSres).

Utilizing the data from 10 rater pairs, rating the 188 barrier items in different dwellings, these computations were conducted as follows. SSR, the sum of squares due to variation across the 10 rater pairs, is:

$$SSR = \sum_{j=1}^{10}\sum_{k=1}^{188}(\overline{A}_{j.} - \overline{A}_{t})^2,$$

where $\overline{A}_{j.}$ denotes the mean of agreement indices of rater pair j over the presence of all 188 barrier items k, and $\overline{A}_{t}$ denotes the total mean of all agreement indices from all rater pairs over all barriers. SSI, the sum of squares due to variation across the 188 barrier items, is:

$$SSI = \sum_{j=1}^{10}\sum_{k=1}^{188}(\overline{A}_{.k} - \overline{A}_{t})^2,$$

where $\overline{A}_{.k}$ denotes the mean of agreement indices for barrier item k over all 10 rater pairs j. SST, the total sum of squares is:

$$SST = \sum_{j=1}^{10}\sum_{k=1}^{188}(A_{jk} - \overline{A}_{t})^2,$$

where A_{jk} denotes the agreement indices of rater pair j on the presence of all barrier items k. Thus, the residual sum of squares of agreement variation is:

$$SSres = SST - (SSR + SSI)$$

The sum of squares can be "translated" into R-square values, indicating the relative share of variance in the agreement indices to be attributed to the respective source of variance. The R-squares indicating the relative share of agreement variation attributable to the rater pair

(R_R^2), the item (R_I^2), and residual impacts (R_{res}^2) are then computed respectively as follows:

$$R_R^2 = \frac{SSR}{SST},\ R_I^2 = \frac{SSI}{SST},\ R_{res}^2 = \frac{SSres}{SST}$$

Thus, we computed the sum of squares of the agreement variation due to variation of rater pairs and due to variation of barrier items. With respect to the components impacting on agreement, we considered theoretically, these sums of squares may be attributed to the respective component. That is, SSR covers variation in agreement produced by varying characteristics of the rater pairs, whereas SSI reveals variation caused by varying characteristics of the barrier items. The residual share of variation in the raters × items design thus covers variation due to all other sources of impact.

Predictors of agreement variation analysis

In our raters × items design, the indices of agreement on the 188 barrier items were nested within 10 rater pairs, forming a hierarchical multilevel data structure. Thus, multilevel regression analysis [24] is well suited for analyzing the impact of hypothetically assumed predictors that may impact on inter-rater agreement. Basically, the multilevel model for the raters × items data implies a decomposition of variance "within raters" (level 1) and "between raters" (level 2), and predictors varying within or between the rater pairs may be specified in the regression equation. In particular, we run a 2-level random intercept models as follows:

$$A_{jk} = \beta_{0j} + \sum_m \beta_m X_{mjk} + \varepsilon_{jk},$$

where A_{jk} is agreement index (i.e. P_o or κ) for rater pair j rating barrier item k; β_{0j} is the regression intercept specific for raters j and β_m is the regression coefficient for predictor m;

X_{mjk} is the value of level-1 predictor m for raters j and barrier item k; ε_{jk} is the level-1 residual value. Thus, for reasons of model parsimony and as our focus was on the fixed effects of the predictors only, we did not model random slopes with respect to the level-1 predictors. The level-2 equation may then be written:

$$\beta_{0j} = \gamma_0 + \sum_n \gamma_n X_{nj} + \upsilon_j,$$

where γ_0 is the regression intercept for the level-2 equation predicting β_{0j}; γ_n is the regression coefficient for level-2 predictor X_{nj} predicting β_{0j}; υ_j is the level-2 residual in predicting β_{0j}. Thus, the impact of level-1 predictors X_{mjk}, varying across the barrier items, could be analyzed as well as of level-2 predictors X_{nj} varying only on the raters-level. X_{mjk} could, for example, be a characteristic of the barrier items to be rated such as its estimated prevalence. P-values<0.05 were considered statistically significant. Multilevel models were run by use of SAS software, PROC MIXED, by choice of options as recommended for multilevel regression models [25]. That is, in particular modelling an "unstructured" covariance structure and running the restricted maximum likelihood estimation (for details see SAS Institute Inc., Cary, NC USA).

RESULTS

The level of agreement

Analysing the mean of all 1,880 observations, the two indices indicated widely different levels of agreement. The mean level of P_o was 0.83 (SD 0.20), whereas the mean level of κ was 0.35 (SD 0.40). Further differences were unfolded when analysing the mean level of agreement by characteristics of the components. When both raters in a rater pair were familiar

with the use of standardized instruments, the average level of agreement was considerably higher, compared to when only one or none of the raters had such experience. However, rater pairs with both raters having housing adaptation experience had slightly lower level of agreement compared to the other rater pairs. Barriers assessed by means of evaluative judgements generated lower levels of agreement, both compared to barriers assessed by means of measurements or as obvious by observation. Agreement level measured by means of P_o tended to be lower by barrier prevalence. As regards κ, it was highest when prevalence was 41-60%. For detailed figures, see Table 2.

Table 2. Mean levels of agreement, by rater and item characteristics.

	Agreement index *Mean (SD)*	
Rater and item characteristic	**Observed agreement** P_o *N=1,880*	**Kappa** κ *N=1,402* [a]
Housing adaptation experience (raters)		
Both raters	0.81 (0.23)	0.31 (0.39)
Only one/none	0.85 (0.19)	0.38 (0.41)
Familiarity with standardized instruments (raters)		
Both raters	0.92 (0.12)	0.57 (0.40)
Only one/none	0.81 (0.22)	0.30 (0.39)
Barrier assessment type (items):		
Measurable	0.84 (0.17)	0.42 (0.40)
Obvious by observation	0.87 (0.21)	0.42 (0.44)
Evaluable	0.80 (0.23)	0.23 (0.35)
Barrier prevalence estimate (items)[b]		
0 - 20 %	0.91 (0.14)	0.37 (0.42)
21 - 40 %	0.79 (0.22)	0.32 (0.39)
41 - 60 %	0.77 (0.20)	0.41 (0.40)
61 - 80 %	0.76 (0.23)	0.35 (0.39)
81-100 %	0.76 (0.29)	0.26 (0.38)
Total	0.83 (0.20)	0.35 (0.40)

[a] Kappa has missing values due to division by zero, i.e. agreement index is undefinable.
[b] Barrier prevalence is estimated as the occurrence in the ENABLE-AGE sample.

Applying the barrier prevalence estimate, 21 items with prevalence in the range from 41-60% were considered the most prevalence balanced. Dividing these items by type of assessment and by raters' familiarity with using standardized assessment instruments revealed apparent differences in the level of agreement, as shown in Table 3. For the rater pairs where both raters had a recorded familiarity of using standardized assessment instruments, the mean levels of agreement were consistently higher for all assessment types, compared to the rater pairs where only one or none of the raters had such familiarity. The highest level of agreement was achieved for items assessed by means of measurement by the raters familiar with the use of standardized instruments ($P_o = 0.93$, $\kappa = 0.80$). In contrast the lowest level of agreement was obtained by the rater pairs without such experience when assessing items by means of evaluative judgement ($P_o = 0.68$, $\kappa = 0.11$).

Table 3. The most prevalence balanced environmental barrier items (n=21), prevalence estimate 41% - 60%.

Item description[a]	**Assessment type**[b]	**Prevalence estimate**[c] **%**		**Agreement index**			
				Observed agreement P_o		**Kappa** κ	
				Both raters are familiar with using standardized instruments			
				Yes	*No*	*Yes*	*No*
Stairs/thresholds/differences in level between rooms/floor	M	42		0.93	0.88	0.80	0.64
Narrow paths (outdoor)	M	44		0.96	0.71	0.93	0.27
No level area in front of entrance doors	M	44		0.89	0.79	0.79	0.51
Toilet with standard height or lower	M	45		0.93	0.76	0.85	0.32
Elevated toilet or higher	M	45		0.96	0.72	0.93	0.37
Narrow passages/corridors in relation to fixtures/design	M	47		0.86	0.79	0.44	0.49
Insufficient manoeuvering areas (kitchen)	M	47		0.93	0.77	0.85	0.51
No surface at a height suitable for sitting work (kitchen)	M(O)	54		0.96	0.79	0.88	0.53
Storage areas can only be reached via stairs/threshold	M	54		0.89	0.71	0.65	0.11
Insufficient manoeuvering space (refuse bin and/or letterbox)	M	55		0.89	0.66	0.73	0.29
Wallmounted cupboards & shelves placed extremely high (kitchen)	M	55		1.00	0.71	1.00	0.36
Refuse bin and/or letterbox difficult to reach	M	56		0.93	0.79	0.79	0.39
			Mean (SD)	*0.93 (0.04)*	*0.76 (0.06)*	*0.80 (0.15)*	*0.40 (0.14)*
Doors that do not stay in open position/close quickly	O(E)	44		0.93	0.77	0.83	0.30
Doors that cannot be fastened in open position	O	46		0.93	0.58	0.86	-0.01
No grab bars at shower/bath and/or toilet	O	48		0.89	0.81	0.79	0.58
No place to sit in shower/bath	O	57		0.86	0.83	0.65	0.42
			Mean (SD)	*0.90 (0.03)*	*0.75 (0.11)*	*0.78 (0.09)*	*0.32 (0.25)*
Inappropriate design of wardrobes/clothes cupboards	E	45		0.79	0.60	0.05	-0.06
Very small controls (kitchen)	E	45		0.86	0.76	0.54	0.21
Insufficient/inappropriately designed lighting (kitchen)	E	47		0.89	0.78	0.73	0.36
Very small controls (other than kitchen or hygiene area)	E	48		0.79	0.80	0.60	0.01
Use requires intact fine motor control (hygiene area)	E	54		0.79	0.44	0.61	0.05
			Mean (SD)	*0.82 (0.05)*	*0.68 (0.15)*	*0.51 (0.26)*	*0.11 (0.17)*

[a] For a complete item list and description of the Housing Enabler instrument, see [17].
[b] M = Measurable, O = Obvious by observation, E = Evaluable. Two items were mixed in type of assessment, with supplementary type in parenthesis.
[c] Barrier prevalence is estimated as the occurrence in the ENABLE-AGE sample.

Shares of agreement variation

As shown in Table 4, the two agreement indices showed similar patterns in variance shares. The raters accounted for 6-11% of the variance, the items accounted for 32-33% of the variance and the residual accounted for 57-60% of the variance. That is, varying characteristics of the raters and/or the items altogether explained about 40% of the variation in P_o and about 43% of κ variation.

Table 4. Shares of agreement variation accounted for by the components of agreement.

	Agreement index[a] *(% of total variance)*	
Component of agreement	**Observed agreement** P_o *N=1,880*	**Kappa** κ *N=1,402* [b]
Raters	6.4	11.0
Items	33.4	31.7
Residual	60.2	57.3

[a] The agreement values' total sum of squares (R^2) decomposed into relative shares accounted for.
[b] Kappa has missing values due to division by zero, i.e. agreement index is undefinable.

Predictors of agreement variation

Results for the multilevel regression model are shown in Table 5. In terms of statistical significance, for both agreement indices item assessment type, prevalence estimate and raters' familiarity with standardized assessment instruments appeared as substantial predictors, whereas raters' housing adaptation experience did not. With respect to the sign of the effects, disagreement increases if the barriers are assessed by means of evaluated judgement and if one or both raters are not familiar with standardized instruments. Regarding the effect of the prevalence estimate, the findings indicate that both P_o and κ values tend to decrease with higher prevalence. As could be seen from the R^2 values printed in Table 4, the four predictors altogether account for some substantial, but not too large share of agreement variation. As

these R^2 values indicate variance contribution from particular characteristics of items and raters, it may be taken as further explanation of the 40-43% of the non-residual variance due to the general variance decomposition reported above and in Table 4.

Table 5. Predictors of agreement variation.

	Agreement index[a]			
Rater and item characteristic	**Observed agreement** P_o *N=1,880*		**Kappa** κ *N=1,402* [b]	
	Est.[e]	*P*	Est.[e]	*P*
Housing adaptation experience (raters)[c]	-0.024	0.444	-0.048	0.529
Familiarity with standardized instruments (raters)[c]	0.107	**0.009**	0.270	**0.007**
Barrier assessment type (items):		**<0.0001**		**<0.0001**
- evaluable vs. obvious	-0.094	**<0.0001**	-0.205	**<0.0001**
- measurable vs. obvious	-0.022	0.060	-0.010	0.717
Barrier prevalence estimate (items)[d]	-0.258	**<0.0001**	-0.099	**0.010**
Level-1 R^2	0.16		0.12	

[a] The agreement indices are treated as dependent variables in the model.
[b] Kappa has missing values due to division by zero, i.e. agreement index is undefinable.
[c] Dichotomized: 0=”Only one/none of the raters experienced/familiar”, 1=”Both raters experienced/familiar”.
[d] Barrier prevalence is estimated as the occurrence in the ENABLE-AGE sample.
[e] Estimated regression coefficient (fixed effect).

DISCUSSION

The main contribution of this study is a proposed multi-component approach for in-depth examination of inter-rater agreement. Ultimately it is intended as a strategy for identification of means to improve instrument reliability, particularly within the health sciences. The proposed approach evolved from a conceptual analysis, decomposing agreement into three main components: raters, items and contexts. Based on this conceptual analysis, we defined a statistical formula for the calculation of relative shares of agreement variation, making it

possible to disentangle the contribution of each component to the total variance in data. As demonstrated with empirical data, applying this formula serves to unfold the complex phenomenon of inter-rater agreement. Proceeding with multilevel regression analysis gives an even deeper understanding, as significant predictors of agreement variation can thus be identified. To the best of our knowledge, this study represents a new analytic approach to research in the field of inter-rater agreement.

Many inter-rater agreement studies find it sufficient to analyse the level of agreement as an indicator of instrument reliability (see e.g. [26-27]). However, a crucial weakness of the conventional approach is the limited inference of the results that can be based on such analysis [5]. Calling Generalizability theory to mind [11,15], our analytical strategy aims to promote knowledge under what conditions inferences can be made. As we consider agreement as something relative to the aspects of the components, this relative nature is likely to be reflected as a variation in agreement, to the extent that the aspects of the components vary between studies. Without knowing what impacts on the variation in agreement in the first place, it is hazardous to predict a similar outcome in another study, where aspects of the components may differ. In addition to examining the level of agreement, we therefore propose to analyse the shares of variation accounted for by the components, which provides guidance as to the domain of where to find the main sources of disagreement. In the current study we found that even though the level of agreement differed between P_o and κ, the shares of variation were more or less comparable. The residual accounted for the major part of the agreement variation, suggesting that essential sources of disagreement presumably are to be found among contextual characteristics. Although the residual also comprises "conventional" error due to unsystematic impacts, we would argue that most of such unsystematic impacts may be considered to be produced by contextual particularities. An interpretation of the

residual predominance could be that even with “optimal” raters and items, there is a limit as to how much the level of agreement can be improved when contextual circumstances are not sufficiently controlled for. Even so, the result also indicates that the level of agreement can be substantially improved by identifying and counteracting those characteristics of the components tending to negatively influence agreement. To distinguish those characteristics however, we propose to use multilevel regression analysis. With regard to the multitude of potential characteristics of the components which could go into such analysis (see Figure 1), it is equally important to have an open mind when considering relevant characteristics to include in data collection as to have a well thought-out study design which allows for the multilevel data structure necessary for this analysis.

In the current study, we were restricted to two characteristics of the rater component and two of the item component. In hindsight, and in light of the result of the shares of agreement variation analysis, it would have been desirable to have had contextual data systematically collected as well. Keeping in mind that the empirical data we used mainly served an instrumental purpose, inferences made from the results have to take this study limitation into account. Nevertheless, we have demonstrated the potential benefit of this kind of analysis, pointing to rater and item characteristics that should be remedied in order to improve agreement. It was not surprising to find that barrier prevalence estimate, lack of familiarity with using standardized assessment instruments and item ratings depending on evaluative judgements came out as significant predictors. With additional data on other characteristics of the components, this analysis could have been taken even deeper. In particular, the contextual emphasis suggested by our results deserves to be explored by further research.

The sample sizes used in the multilevel regression analysis also need to be considered. With only 10 rater-pairs, which constitute the level-2 sample size for our multilevel model, conventional sample size recommendations [24] such as the 30/30 rule (i.e., 30 level-2 units, each containing 30 level-1 units at least) were not met. In contrast, the level-1 sample size (N=1,880) largely exceeds requirements given in the multilevel modeling literature [24]. Yet, the literature is not definitely conclusive about sample sizes, as respective considerations depend on the type of parameter and statistic potentially affected by low sample size. Recently, Bell et al. [28] examined the performance of two level models under less than ideal conditions, including designs with level-2 sample size of 10. The results suggest that even with small level-2 sample sizes confidence intervals and Type I errors are estimated fairly well and estimates are unbiased. However, the power of the significance tests of the effects of level-2 predictors was substantially decreased. Thus, it should be kept in mind that our design may be underpowered with regard to the impact of rater characteristics.

Approaches which implicitly consider variation of agreement have also been suggested within the conceptual framework of Bayesian statistics, basically employing computations of the posterior probabilities of agreement by combining agreement data with prior information on the distribution of agreement [29]. This implies to model the probability of agreement conditional on, for example, the objects to be rated and/or the raters. Our study did not follow the Bayesian approach to agreement in modeling posterior probabilities, but may be viewed as conceptually related in that we aimed to explore variables hypothetically contributing to the variance of the agreement, hence moderators of agreement probability. With such sources of agreement variation identified, future research may consider their implementation in Bayesian strategies such as the resampling methods to estimate posterior probabilities as proposed by Broemeling [30].

Inter-rater agreement studies are important as means to establish reliable use of assessment instruments. In a welcome addition to the literature on agreement/reliability studies [9], commendable guidelines for reporting such studies were recently proposed. The present study contributes further to this research field, by exploring and proposing new strategies for in-depth examination of agreement data. Using a multi-component strategy, where the different steps complement and strengthen each other, our approach focuses on identifying the most important sources of disagreement as targets for remedying measures. Therefore, for future studies on instruments involving contextualised assessments, we recommend a study design and data collection that enables such an analytical strategy, systematically crossing characteristics of raters, items and contexts. That would enhance the possibilities of detecting weaknesses threatening reliable instrument use, yielding a basis for refinement of the instruments themselves, better rater training and a raised awareness of potential impacts of various contextual circumstances. In conclusion, our recommendations for study design and data analysis have the potential of ultimately improving the reliability of assessment instruments, so important for adequate measures taken, as well as for efficient resource allocation.

REFERENCES

[1] Fayers PM, Machin D. Quality of Life - Assessment, Analysis and Interpretation. John Wiley & Sons, Ltd. New York, USA; 2000.

[2] Kozlowski SWJ, Hattrup K. A disagreement about within-group agreement: disentangling issues of consistency versus consensus. Journal of Applied Psychology 1992;77:161-167.

[3] De Vet HCW, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. Journal of Clinical Epidemiology 2006;59:1033-1039.

[4] Kottner J, Streiner DL, The difference between reliability and agreement. Journal of Clinical Epidemiology 2011;64:701-702.

[5] Streiner DL, Norman GR. Health Measurement Scales. A Practical Guide to Their Development and Use. 4th ed Oxford University Press, Oxford; 2008.

[6] Helle T, Nygren C, Slaug B, Brandt Å, Pikkarainen A, Hansen A-G, Pétersdóttir E, Iwarsson S. The Nordic Housing Enabler: Inter-rater reliability in cross-Nordic occupational therapy practice. Scandinavian Journal of Occupational Therapy 2010;17:258-266.

[7] Iwarsson S, Nygren C, Slaug B. Cross-national and multi professional inter-rater reliability of the Housing Enabler. Scandinavian Journal of Occupational Therapy 2005;12:29-39.

[8] Park J-K, Boyer J, Tessler J, Casey J, Schemm L, Gore R, Punnett L. & Promoting Healthy and Safe Employment (PHASE) in Healthcare Project Team. Inter-rater reliability of PATH observations for assessment of ergonomic risk factors in hospital work. Ergonomics 2009;52:820-829

[9] Kottner J, Audige L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, Roberts C, Shoukri M, Streiner DL. Guideline for reporting reliability and agreement studies (GRRAS) were proposed. Journal of Clinical Epidemiology 2011;64:96-106.

[10] Schwartz CE, Rapkin BD. Reconsidering the psychometrics of quality of life assessment in light of response shift and appraisal. Health and Quality of Life Outcomes 2004;2:16. Bio Med Central, open access: Http://www.hqlo.com/content/2/1/16.

[11] Brennan RL. Generalizability theory. Springer-Verlag, New York; 2001.

[12] Bloch DA, Kraemer HC. 2 × 2 Kappa Coefficients: Measures of agreement or association. Biometrics 1989;45:269-287.

[13] Kraemer HC. Ramifications for a population model for κ as a coefficient of reliability. Psychometrika 1979;44:461-472.

[14] Phillips CD, Patnaik A, Dyer JA, Naiser E, Hawes C, Fournier CJ, Elliott TR. Reliability and the measurement of activity limitations (ADLs) for children with special health care needs (CSHCN) living in the community. Disability and Rehabilitation 2011:1-10.

[15] Cronbach LJ, Gleser GC, Nanda H, Rajaratnam N. The dependability of behavioral measurements: Theory of generalizability of scores and profiles. Wiley, New York; 1972.

[16] Mitchell SK. Interobserver agreement, reliability, and generalizability of data collected in observational studies. Psychological Bulletin 1979;86:376-390.

[17] Iwarsson S, Slaug B. The Housing Enabler: An instrument for assessing and analysing accessibility problems in housing. Nävlinge & Staffanstorp: Veten & Skapen HB & Slaug Data Management; 2001.

[18] Iwarsson S, Wahl H-W, Nygren C, Oswald F, Sixsmith A, Sixsmith J, Széman Z, Tomsone S. Importance of the home environment for healthy aging: Conceptual and methodological background of the European ENABLE-AGE Project. Gerontologist 2007;47:78-84.

[19] Oswald F, Wahl H-W, Schilling O, Nygren C, Fänge A, Sixsmith A, Sixsmith J, Széman Z, Tomsone S, Iwarsson S. Relationships between housing and healthy ageing aspects in very old age: Results from the European ENABLE-AGE Project. Gerontologist 2007;47:96-107.

[20] Hripcsak G, Heitjan DF. Measuring agreement in medical informatics reliability studies. Journal of Biomedical Informatics 2002;35:99-110.

[21] Fleiss JL, Levin B, Paik MC. Statistical Methods for Rates and Proportions, Third Edition, New York: John Wiley & Sons; 2003.

[22] Shoukri MM. Measures of inter-observer agreement. Chapman & Hall/CRC, Washington DC; 2004.

[23] Hohler FK. Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. Journal of Clinical Epidemiology 2000;53:499-503.

[24] Hox JJ. Multilevel analysis. Mahwah: Lawrence Erlbaum; 2002.

[25] Singer JD. Using SAS PROC MIXED to Fit Multilevel Models, Hierarchical Models, and Individual Growth Models. Journal of educational and behavioral statistics. 1998; 23:323-355.

[26] Ghotbi N, Ansari NN, Naghdi S, Hasson S, Jamshidpour B, Amiri S. Inter-rater reliability of the modified Ashworth scale in assessing lower limb muscle spasticity. Brain 2009;23:815-819.

[27] Holmefur M, Krumlinde-Sundholm L, Eliasson A-C. Interrater and Intrarater Reliability of the Assisting Hand Assessment. American Journal of Occupational Therapy 2007;61:79-84.

[28] Bell BA, Morgan GB, Schoeneberger JA, Loudermilk BL, Kromrey JD, Ferron JM. Dancing the sample size limbo with mixed models: How low can you go? SAS Global Forum 2010: Paper 197-2010.

[29] Broemeling LD. Bayesian methods for measures of agreement. Boca Raton (FL): Taylor & Francis; 2009.

[30] Broemeling LD. A Bayesian analysis of inter-rater agreement. Communications in Statistics – Simulation and Computation 2001;30:437-446.

APPENDIX

Agreement indices

Rater A's assessment	Rater B's assessment		Total
	Presence	Absence	
Presence	*a*	*b*	*a + b*
Absence	*c*	*d*	*c + d*
Total	*a + c*	*b + d*	*a + b + c + d*

Observed agreement (P_o) is the proportion of cases for which the raters give the same response. If there are two raters and responses are dichotomous, like in this study, then P_o is the sum along the diagonal divided by the total number of cases given.

The formula for P_o is thus:

$P_o = (a + d) / (a + b + c + d)$

Kappa (κ) calculates the degree of agreement that exceeds the degree of agreement that is expected by chance. The agreement expected by chance is calculated in the same way as the observed agreement, except that the observed values in the cells are replaced with their expected values P_e, which are based on the observed proportion of responses in each category. The expected value in cell a, $P_e[a]$, is $(a + b)(a + c) / (a + b + c + d)$ and the expected value in cell d, $P_e[d]$, is $(b + d)(c + d) / (a + b + c + d)$. The agreement expected by chance is $P_e = (P_e[a] + P_e[d]) / (a + b + c + d)$.

The formula for κ is thus:

$\kappa = (P_o - P_e) / (1 - P_e)$