doi: 10.1111/j.1369-7625.2011.00689.x

Reliability and validity of the German version of the OPTION scale

Oliver Hirsch PhD,* Heidemarie Keller PhD,* Meike Müller-Engelmann PhD,* Monika Heinzel Gutenbrunner PhD,† Tanja Krones MD‡ and Norbert Donner-Banzhoff MD, MHSc§

*Research Fellow, Department of General Practice/Family Medicine, Philipps University Marburg, †Senior Research Fellow, Department of Child and Adolescent Psychiatry and Psychotherapy, Philipps University Marburg, ‡Senior Research Fellow, Department of General Practice/Family Medicine, Philipps University Marburg and §Professor of Primary Health Care, Department of General Practice/Family Medicine, Philipps University Marburg, Germany

Correspondence
Dr Oliver Hirsch
Department of General Practice/Family Medicine
Philipps University Marburg
D-35032 Marburg
Germany
E-mail oliver.hirsch@staff.uni-marburg.de

Accepted for publication
23 March 2011

Keywords: decision making, empirical research, patient participation, physician–patient relations (all MeSH), psychometrics

Abstract

Objective To examine the psychometric properties of the German version of the 'observing patient involvement' scale (OPTION) by analysing video recordings of primary care consultations dealing with counselling in cardiovascular prevention.

Design Cross-sectional assessment of physician–patient interaction by two rater pairs and two experts in shared decision making (SDM).

Setting Primary care.

Participants Fifteen general practitioners provided 40 videographed consultations.

Measurements Video ratings using the OPTION instrument.

Results Mean differences on item level between the four raters were quite large. Most items were skewed towards minimal levels of shared decision making. Measures of inter-rater association showed low to moderate associations on item level and high associations on total score level. Cronbach-α of the whole scale based on the data of all four raters is 0.90 and therefore on a high level. An oblique factor analysis revealed two factors, but both factors were highly correlated so we can confirm a one-dimensional structure of the instrument. ROC analyses between the rater total scores and dichotomized expert ratings (SDM yes/no) revealed a good discriminability of the OPTION total score. Physicians with more expertise in shared decision making received higher OPTION ratings.

Conclusions The German version of the OPTION scale is reliable at total score level. Some items need further revision in the direction of more concrete, observable behaviour. We were only able to perform a quasi-validation of the scale. Validity issues need further research efforts.

Introduction

Patients prefer to be involved in decisions about their medical care as studies conducted in a variety of settings have shown.[1,2] Patient involvement still needs to be measured to determine progress and provide individual feedback.[3,4] Instruments for measuring patient involvement are few as revealed in the relevant literature.[5] Although some instruments include some components of patient involvement,[6–12] they were found to be insufficiently precise to accurately measure this aspect of communication in patient–clinician interactions. The need to investigate the validation of instruments said to measure shared decision making (SDM) is emphasized. Until now, only indirect validity parameters have been used.[13,14]

Against this background, Elwyn *et al.* have developed a scale called 'observing patient involvement' (OPTION), which is supposed to assess the extent to which clinicians involve patients in decisions across a range of situations in clinical practice. The first version of the OPTION scale used attitude scaling on a five-point scale. Its psychometric characteristics were evaluated by two raters rating 186 audiotaped consultations.[10] The mean intraclass correlation coefficient was 0.62, the mean kappa value 0.71, and Cronbach's α was 0.79. Results for the single items were heterogenous, especially low agreement was found on item 9 (clinician provides opportunities to ask questions). A confirmatory factor analysis revealed that the instrument has a one-dimensional structure.

The revised version of the scale uses a magnitude (numerical rating) instead of an attitude scale (verbal rating).[15] The same audio-taped consultations were used to examine its psychometric properties. Factor analysis again confirmed a one-dimensional structure that was interpreted in the way that SDM might be a homogenous construct. The intraclass correlation coefficient for the total score was 0.77. On item level, there was moderate variability between the raters. Kappa scores on item level ranged from 0.45 to 0.98, and intra-class coefficients (ICCs) ranged from 0.11 to 0.98. The intra-rater ICC for the total score was 0.53. Elwyn *et al.* conclude that the OPTION instrument is not reliable at the individual item level but on the total score level.

The reliability of the Italian version of the OPTION scale was assessed by two raters rating thirty consultation transcripts.[16] Weighted kappa values ranged from 0.29 to 0.73, the intraclass correlation coefficients for the total score at test and retest were 0.85 and 0.81, and Cronbach's α was 0.82. The distributions of the single items were skewed with the majority between 0 (behaviour absent) and 2 (minimum skill level).

A construct validity study was carried out by Siriwardena *et al.* Candidates passing the 'sharing management options' in the examination for membership of the Royal College of General Practitioners (MRCGP) had higher OPTION scores than those who failed.[17]

The German version of the OPTION scale was used in the study of Loh *et al.*[18] in the context of depression. Two raters rated 20 consultations with depressive patients in primary care. Apart from an inter-rater concordance of 67%, a Kappa coefficient 0.5 and an ICC of 0.7, no further psychometric characteristics were reported. Consequently, there is a lack of such data concerning the German version of the OPTION scale, and ours is the first study to examine the instrument in detail. The primary aim of our study was to measure inter-rater agreement and inter-rater associations as measures of reliability. Regarding validity, we aimed at examining the factorial structure of the instrument, and we intended to compare the raters' evaluations of the OPTION scale to expert ratings whether shared decision making took place or not, which we regarded as a quasi-validation. This study is part of an extensive phase 4 study investigating patient participation in the SDM process in cardiovascular prevention. It was approved by the local ethics committee of the Department of Medicine at the University of Marburg. Informed consent had been obtained from participating general practitioners (GPs) and patients.

Methods

Sample

All 91 GPs (44 in the intervention and 47 in the control group) of a previous cluster randomized trial were asked to participate in this study.[19] Fifteen GPs (10 men, 5 women) agreed to participate and were asked to recruit three patients each in whom discussion of cardiovascular risk and of preventive measures seemed indicated. GPs and patients agreed to having their consultations videotaped. This resulted in a total of 40 videotaped consultations.

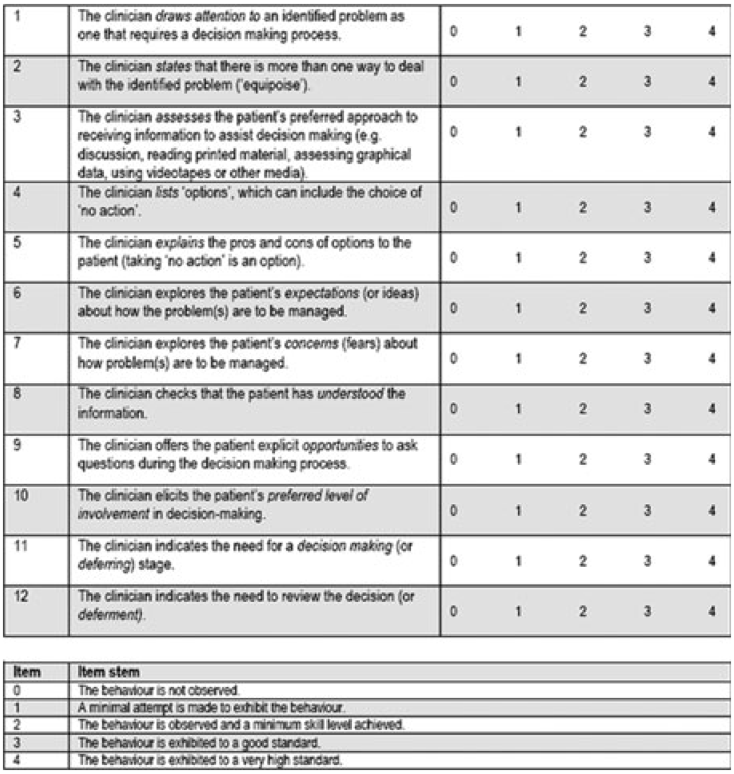
Eight GPs had taken part in a preceding randomized controlled trial where they have received educational training in SDM.[20] Thus, we consider them to be physicians with a certain expertise in SDM.

The instrument

The original OPTION scale consists of a set of competences[21] that include problem definition, explaining legitimate choices, portraying options and communicating risk and conducting the decision process or its deferment. The instrument is a 12-item five-point scale to assess the presence and characteristics of the clinician's communication behaviour (competence). Scaling corresponds to the observed behaviour (0 = not observed, 1 = minimal attempt, 2 = minimal skill level, 3 = good standard, 4 = high standard). A comprehensive account of the four-stage translation process into German is given by Elwyn *et al.*[22] Figure 1 depicts the English version of the OPTION scale.

Procedure

Our OPTION ratings were based on video recordings of consultations and were performed by four experienced raters. All four raters are research fellows with knowledge of the principles of SDM. They were divided into two rater pairs (rater 1 vs. rater 3 and rater 2 vs. rater 4). We conducted an extensive training of the raters, and each rater pair performed a calibration



1	The clinician *draws attention to* an identified problem as one that requires a decision making process.	0	1	2	3	4
2	The clinician *states* that there is more than one way to deal with the identified problem ('equipoise').	0	1	2	3	4
3	The clinician *assesses* the patient's preferred approach to receiving information to assist decision making (e.g. discussion, reading printed material, assessing graphical data, using videotapes or other media).	0	1	2	3	4
4	The clinician *lists* 'options', which can include the choice of 'no action'.	0	1	2	3	4
5	The clinician *explains* the pros and cons of options to the patient (taking 'no action' is an option).	0	1	2	3	4
6	The clinician explores the patient's *expectations* (or ideas) about how the problem(s) are to be managed.	0	1	2	3	4
7	The clinician explores the patient's *concerns* (fears) about how problem(s) are to be managed.	0	1	2	3	4
8	The clinician checks that the patient has *understood* the information.	0	1	2	3	4
9	The clinician offers the patient explicit *opportunities* to ask questions during the decision making process.	0	1	2	3	4
10	The clinician elicits the patient's *preferred level of involvement* in decision-making.	0	1	2	3	4
11	The clinician indicates the need for a *decision making* (or *deferring*) stage.	0	1	2	3	4
12	The clinician indicates the need to review the decision (or *deferment*).	0	1	2	3	4

Item	Item stem
0	The behaviour is not observed.
1	A minimal attempt is made to exhibit the behaviour.
2	The behaviour is observed and a minimum skill level achieved.
3	The behaviour is exhibited to a good standard.
4	The behaviour is exhibited to a very high standard.

Figure 1 English version of the observing patient involvement scale.

session to reach a consensus about their general rating performance. All raters rated the 40 videotaped consultations independently in random order. There was a discourse in the training phase but not in the phase of rating. These data were then compared with a reference standard whether SDM took place (yes/no) rated by two experts of the shared decision-making field associated with our group. In three cases, they had to reach a consensus because of disagreement. Global ratings can be regarded as an effort for quasi-validation of a complex construct and cannot be seen as a gold standard.

Additionally, at the end of the rating of each video, the four raters also provided a global rating whether SDM took place or not.

Statistical methodology

Reliability

Agreement between rater pairs on item level was evaluated by the Wilcoxon test for dependent data. It is recommended that a meaningful difference between two raters occurs at a *P*-value of 0.25 or smaller to reduce the probability of an β error. [23,24] We further calculated the effect size d to assess the magnitude of the mean difference.[25]

As there are no gold standards for the measurement of inter-rater reliability, it is highly recommended to use several measures.[23]

Associations between the two rater pairs on item level were assessed by Spearman correlation coefficients, weighted kappas and intraclass correlation coefficients.[23,26,27] Unadjusted intraclass correlations above 0.40, 0.60 and 0.80 were considered to reflect fair, moderate and substantial agreement.[28] Weighted kappas smaller than 0.40 were taken as an indicator of poor agreement, between 0.41 and 0.59 they signal moderate agreement, between 0.60 and 0.74 they are interpreted as good agreement, and kappa scores larger than 0.74 show very good agreement.[29]

The association of the total scores between the raters was examined by Pearson correlation coefficients. Point–biserial correlation coefficients were calculated between the sum scores of the four raters and their respective dichotomized SDM ratings.[26]

The internal consistency of the whole scale was inspected with Cronbach-α. The previously mentioned coefficients are acceptable at a value of 0.7 and larger.[23] The minimum required sample size in our study was 26 videos while we hoped for an inter-rater reliability of 0.8. Details on sample size calculations for reliability studies are provided by Walter *et al.*[30]

Validity

We used the data of all raters to examine the structure of the scale by principal components analysis with the oblique Promax rotation to allow factors to be correlated. The Kaiser–Meyer–Olkin criterion and the measure of sampling adequacy were used to judge the quality of the factor analytic solution.[31,32] Our hypothesis was that we are able to replicate the one-dimensional structure reported in the literature.

Point–biserial correlation coefficients were calculated between the total OPTION scores of the four raters and the dichotomized SDM expert consensus ratings.[26]

We plotted the total OPTION scores of all four raters against the respective SDM consensus ratings of the experts using ROC analysis to search for a cut-off point that differentiates between 'SDM: yes' and 'SDM: no'.[33]

We further explored differences in total OPTION scores regarding patients' age, gender, education and cardiovascular risk and regarding physicians' expertise in shared decision making by comparing mean differences with *t*-tests and analysis of variance (ANOVA). Calculations were carried out with PASW Statistics 18 (SPSS Inc., Chicago, IL, USA) and MEDCALC 11.2.1.0 (MedCalc Software, Mariakerke, Belgium).

Results

Study sample

Participating GPs were 5 women and 10 men whose age ranged from 44 to 56 years. Regarding age, our study sample corresponds to the age distribution of practicing physicians in Germany

Table 1 Means and standard deviations of the single items and the total score of the OPTION scale listed separately for the four raters (*n* = 40 per rater on each item)

	Rater 1	Rater 2	Rater 3	Rater 4
Item 1	3.74 (0.56)	2.17 (1.25)	2.70 (0.98)	1.79 (0.79)
Item 2	3.16 (1.21)	1.55 (1.36)	2.65 (1.43)	1.58 (1.02)
Item 3	0.22 (0.94)	0.00 (0.00)	0.35 (0.81)	0.00 (0.00)
Item 4	2.74 (1.28)	1.80 (1.36)	2.40 (1.23)	2.21 (1.23)
Item 5	2.21 (1.44)	0.80 (1.06)	1.68 (1.29)	1.26 (0.99)
Item 6	1.89 (1.45)	0.15 (0.37)	1.25 (0.91)	1.00 (1.00)
Item 7	1.74 (1.33)	0.45 (0.83)	1.45 (0.95)	1.11 (1.05)
Item 8	2.29 (1.26)	1.84 (1.77)	1.85 (0.99)	1.16 (0.69)
Item 9	3.68 (0.48)	3.10 (1.12)	2.65 (0.99)	1.47 (1.17)
Item 10	1.06 (1.09)	0.35 (0.88)	1.10 (0.97)	0.37 (0.50)
Item 11	1.47 (1.58)	0.20 (0.41)	2.05 (1.23)	1.63 (0.83)
Item 12	1.95 (1.68)	0.35 (0.81)	1.85 (1.31)	1.68 (1.00)
Total score	24.64 (11.57)	12.50 (7.37)	21.26 (9.51)	15.26 (7.84)

(43% > 54 years, 12% > 60 years, mean age 51.3 years) as published by the National Association of Statutory Health Insurance Physicians in 2009.

From the 45 videos (14 male and 26 female patients), five had to be excluded because during the consultation, other problems different than cardiovascular risk were discussed.

Descriptive statistics

Table 1 depicts descriptive statistics of the four raters on item and total score level. We calculated the raw total score without further transformation.

One can see in Table 1 that the mean differences on item level between the four raters are quite large. Most items are skewed towards minimal levels of SDM. The calibration process of the rater pair 1 and 3 was different from the one of rater pair 2 and 4. This is especially obvious in the difference of the total scores. Item 3 (clinician assesses patient's preferred approach to receiving information) was almost never observed by all four raters.

Reliability

Inter-rater association

Table 2 *P*-values of the Wilcoxon test between the two rater pairs

	Wilcoxon test (*P*-value) rater 1 vs. rater 3	Wilcoxon test (*P*-value) rater 2 vs. rater 4
Item 1	< **0.001**	**0.19**
Item 2	**0.05**	0.81
Item 3	0.26	1.00
Item 4	0.27	**0.22**
Item 5	0.35	**0.09**
Îtem 6	**0.10**	**0.003**
Item 7	0.35	**0.005**
Item 8	**0.12**	**0.06**
Item 9	**0.01**	**0.001**
Item 10	0.73	0.85
Item 11	**0.07**	< **0.001**
Item 12	0.98	< **0.001**

Significant values according to the recommendations by Wirtz and Caspar are printed in bold.

Table 2 depicts the results between the two rater pairs regarding their associations on the items of the OPTION scale. Table 2 shows that the distributions of the ratings especially differ between raters 2 and 4 with significant differences except for items 2, 3 and 10. The agreement between raters 1 and 3 is higher with non-significant differences on items 3, 4, 5, 7, 10 and 12.

Effect sizes between rater 1 and rater 3 range between d = 0.04 and 1.55. Between rater 2 and rater 4, effect sizes vary between d = 0.00 and 2.27. Associations between the two rater pairs are shown in Tables 3 and 4.

In Table 3, Spearman correlation coefficients between raters 1 and 3 reflect moderate associations on items 1, 2, 4 and 5. Only low associations were found on items 6 (clinician explores patient's expectations), 7 (clinician explores

Table 3 Measures of association between rater 1 and rater 3 on single items of the OPTION scale (n = 40 videos)

	Spearman	Intraclass correlation	Weighted kappa
Item 1	0.57 (P = 0.01)	0.23 (P = 0.04)	0.14
Item 2	0.76 (P < 0.001)	0.76 (P < 0.001)	0.54
Item 3	0.55 (P = 0.02)	0.76 (P < 0.001)	0.64
Item 4	0.54 (P = 0.02)	0.61 (P = 0.002)	0.34
Item 5	0.48 (P = 0.04)	0.51 (P = 0.012)	0.40
Item 6	0.34 (P = 0.16)	0.29 (P = 0.11)	0.18
Item 7	0.27 (P = 0.26)	0.31 (P = 0.09)	0.14
Item 8	0.54 (P = 0.02)	0.58 (P = 0.005)	0.39
Item 9	−0.45 (P = 0.06)	−0.54 (P = 0.99)	−0.17
Item 10	0.66 (P = 0.004)	0.64 (P = 0.002)	0.52
Item 11	0.60 (P = 0.007)	0.42 (P = 0.03)	0.25
Item 12	0.75 (P < 0.001)	0.71 (P < 0.001)	0.44

Table 4 Measures of association between rater 2 and rater 4 on single items of the OPTION scale (n = 40 videos)

	Spearman	Intraclass correlation	Weighted kappa
Item 1	0.32 (P = 0.22)	0.19 (P = 0.23)	0.07
Item 2	0.66 (P = 0.002)	0.68 (P = 0.001)	0.47
Item 3	*	*	*
Item 4	0.62 (P = 0.004)	0.62 (P = 0.002)	0.40
Item 5	0.69 (P = 0.001)	0.52 (P = 0.008)	0.36
Item 6	0.50 (P = 0.03)	0.07 (P = 0.38)	0.12
Item 7	0.58 (P = 0.009)	0.18 (P = 0.22)	0.15
Item 8	0.28 (P = 0.25)	0.12 (P = 0.31)	0.09
Item 9	0.39 (P = 0.10)	−0.08 (P = 0.63)	0.13
Item 10	−0.05 (P = 0.85)	−0.03 (P = 0.56)	0.05
Item 11	−0.24 (P = 0.32)	−0.45 (P = 0.98)	0.02
Item 12	0.31 (P = 0.19)	−0.06 (P = 0.60)	0.07

*Coefficients for item 3 could not be calculated because of too many '0' values

patient's concerns) and 12 (need to review the decision). Intraclass correlation coefficients signal moderate associations on items 2, 3, 4, 10 and 12. Weighted kappas reveal a good agreement on item 3 and fair agreement on items 2, 5, 10 and 12. It is striking that on item 9 (clinician offers the patient opportunities to ask questions), there is a negative association and agreement between raters 1 and 3.

In Table 4, Spearman correlation coefficients between raters 2 and 4 reflect moderate associations on items 2, 4, 5, 6 and 7. Intraclass correlation coefficients signal moderate associations only on items 2 and 4. Weighted kappas reveal a moderate agreement on items 2 and 4. Low or even negative associations and agreement are found on items 8–12.

The Pearson correlation between the sum scores of rater 1 and rater 3 is 0.68 (P = 0.01) and 0.82 (P < 0.001) between rater 2 and rater 4. They can be considered high.

The point–biserial correlations between the sum scores of the four raters and their respective dichotomized SDM ratings were 0.71, 0.75, 0.79 and 0.81 (all P-values <0.001). This highlights that the raters incorporated their OPTION ratings into their overall decision whether SDM took place or not.

Internal consistency

Cronbach-α of the whole scale based on the data of all four raters is 0.90 and therefore on a high level. The corrected item–total correlations of items 3, 8 and 9 are moderate at around 0.40. The other corrected item–total correlations range from 0.51 to 0.82 and are acceptable.[26,31]

Validity

Factor analysis

In our factor analysis, the Kaiser–Meyer–Olkin criterion with 0.88 is high, and such an analysis is therefore feasible. The measures of sampling adequacy of the single variables range between 0.80 and 0.94 and point in the same direction. Two factors with eigenvalues > 1 were extracted, which together explain 64.5% of the variance (50.3 and 14.2%, respectively). The two factors correlate by r = 0.53. The first factor consists of items 3, 5, 6, 7, 11 and 12, and the second factor comprises items 1, 2, 4, 8, 9 and 10. Cronbach-α of both factors is 0.87, respectively. Correlations of the OPTION scale items with the two extracted factors are depicted in Table 5.

Because of the relatively high correlation between the two factors, cross-correlations of about half of the items are also high.

We calculated point–biserial correlations between the total scores of the four raters and the

Table 5 Correlations of the OPTION scale items with the two extracted factors after Promax rotation (structure matrix)

	Factor 1	Factor 2
Item 1	0.53	**0.81**
Item 2	0.70	**0.82**
Item 3	**0.46**	0.29
Item 4	0.71	**0.73**
Item 5	**0.79**	0.66
Item 6	**0.86**	0.58
Item 7	**0.80**	0.50
Item 8	0.29	**0.75**
Item 9	0.17	**0.78**
Item 10	0.63	**0.64**
Item 11	**0.80**	0.15
Item 12	**0.84**	0.27

The highest loadings of each item are printed in bold.

dichotomous SDM expert consensus ratings of the two experts. The resulting correlations are $r = 0.30$ ($P = 0.07$), $r = 0.32$ ($P = 0.05$), $r = 0.50$ ($P < 0.001$) and $r = 0.51$ ($P < 0.001$).

In a next step, we plotted the sum scores of all four raters against the respective SDM ratings of the experts using ROC analysis. The ROC curve is shown in Figure 2.

The diagonal in Figure 2 represents chance level with an area under the curve (AUC) of 0.50. The AUC in our study is 0.82 ($P < 0.001$; 95% confidence interval: 0.70–0.90). This value is high and shows good discrimination by the OPTION total score.

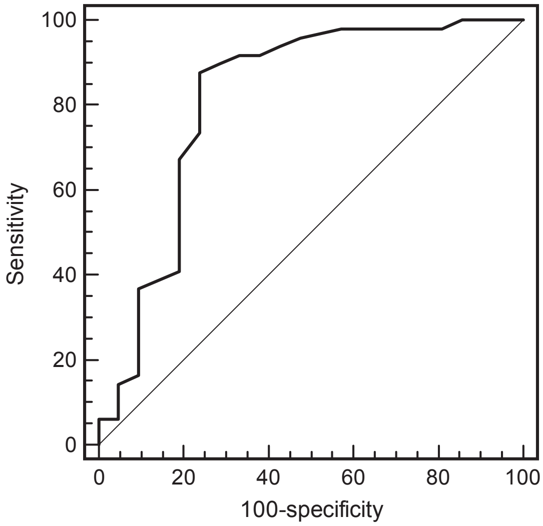


Figure 2 ROC analysis of the sum scores of all four raters and the respective expert shared decision making consensus ratings.

At a cut-off of 12 points (25 points in the scaling with a maximum of 100), the sensitivity is 87.8%, and the specificity is 76.2%. This is the optimum relationship between these two measures in our data.

After a median split of patients' age, we found a significant difference in the total OPTION score (*t*-test, $P = 0.037$). Consultations with patients under 63 years of age were rated lower (mean 16.66, SD 9.58) than consultations with patients equal to or over 63 years of age (mean 21.36, SD 10.19). This results in a medium effect size of $d = 0.48$.[34] In other patient characteristics (gender, education and cardiovascular risk), no significant differences were found. Physicians with more expertise, as measured by the participation in special courses within our randomized controlled trial,[19] received higher OPTION ratings (mean 23.64, SD 8.16) than those with less expertise (mean 15.43, SD 10.14). This results in a large effect size of $d = 0.90$.[34]

Discussion

We were able to show that the German version of the OPTION scale is reliable at the total score level. Statistical quality criteria on item level are heterogenous in that some items had low levels of agreement and association between raters while these values in other items were in an acceptable range.

Some limitations of our study have to be discussed. In contrast to the other reliability studies, we only had one index problem, namely cardiovascular prevention that limits content variation. We used videotaped consultations so that the raters had additional information in the form of non-verbal behaviour (e.g. gestures, facial expressions), which might have led to increased variance between the raters. The data were partly dependent because each physician provided up to three consultations. The consultation times in the Italian and UK reliability studies are shorter than in our study, which also might have caused greater variability between the raters. We were only able to apply quasi-validation criteria. The question of the validity of SDM measures could not be resolved by this study. Descriptive statistics show the rater dependency of the OPTION scale because there were larger variations in mean values on several items between the two rater pairs. Besides the common rater training, the two rater pairs underwent different calibration processes resulting in larger differences on item and total score level.

This finding is verified when examining the differences within each rater pair. The best agreement was found on items with low means. This shows that agreement is highest on the items where SDM not or rarely took place.

The coefficients of association within the two rater pairs are acceptable as long as the behaviour of the clinician is on a more concrete level. They are lower on items where there is more room for interpretation (e.g. explores the patient's expectations; explores the patient's concerns; offers explicit opportunities to ask questions; elicits the patient's preferred level of involvement). Therefore, there might be a need for a revision of such items to a more concrete level so that raters have a common understanding of what explicit behaviour should be measured. Weighted kappas were more often divergent from Spearman coefficients and ICCs. This might have happened because of skewed distributions, which is a known phenomenon.[35,36] In our study, we observed larger ranges of weighted kappa values and ICC values on item level than in the Italian and UK versions.

We also found high associations on the level of the total score. This confirms the results of the other reliability studies[10,15,16,22] and leads us to the conclusion that the total score is an acceptable parameter for further use.

The high associations between the dichotomized SDM ratings of the four raters and their respective total OPTION scores highlight that the raters incorporated their OPTION ratings into their overall decision whether SDM took place or not. The raters consistently connected their OPTION ratings with their concept of SDM, which can be regarded as an indication for intra-rater reliability.

The internal consistency of the scale measured by Cronbach-α is high, which also confirms the findings of the other studies.[10,15,16,22]

Like Elwyn *et al.*,[15] we calculated an exploratory factor analysis with oblique rotation which also resulted in a two-factor solution. The correlation between the two factors in our solution is $r = 0.53$. A one-factor solution in our study had a higher proportion of explained variance than reported by Elwyn *et al.* (50.3 vs. 28%). Pett *et al.*[32] recommend to drop one or more factors when such a large correlation occurs. We therefore can also confirm the one-dimensional structure of the OPTION scale.

The correlations between the total scores of the raters and the dichotomized SDM expert consensus ratings are satisfactorily and lead to the conclusion that the OPTION total score can be used for further analyses. The good discriminability of the OPTION total score was verified by the results of our ROC analyses. The AUC value was high and confirmed that the total score is able to achieve a differentiation between expert ratings of SDM taking place or not even though the cut-off point is quite low. Whether this is an indicator of validity is controversial because this can also be interpreted as a parameter of inter-rater reliability between the raters in our study and the experts. Like Elwyn *et al.*,[15] we also think that SDM could be present in consultations with low OPTION scores although doctors fail to exhibit several basic competencies necessary for SDM. On the other hand, it can also be possible that the OPTION scale measures physician behaviour that is appropriate for patient involvement, which not necessarily means that the OPTION scale measures SDM. Weiss and Peters[37] compared the OPTION scale to the Informed Decision Making Instrument and found an unacceptable level of agreement between the two instruments. This was surprising because both were considered to measure common aspects of SDM. Obviously, SDM is a complex, multidimensional construct that not only consists of distinct stages but also requires special basic interactive and communicative skills of the physician.

The finding that physicians with more expertise in shared decision making received higher OPTION ratings is another aspect that supports the suitability of the OPTION total score. Our result showing higher OPTION scores in elderly patients contradicts other research in this area[38] and needs further exploration.

Conclusions

The German version of the OPTION scale is reliable at total score level. Some items need further revision in the direction of more concrete, observable behaviour. At the present stage, this limits the applicability of the scale. On the other hand, a detailed manual with illustrative examples might also improve statistical quality criteria on item level. We were only able to perform a quasi-validation of the scale. Validity issues need further research efforts. The observer version of the scale is physician centred and neglects the activity and a possible self-involvement of the patient. Therefore, we modified the scoring of the OPTION scale. The results will be presented in a different article. Melbourne *et al.*[39] recently described the development process of a dyadic OPTION scale. This should be one of the topics of future research regarding the OPTION scale to further refine it.

Acknowledgements

Grateful thanks to Andreas Loh who provided training in rating the OPTION scale. We acknowledge with thanks the assistance of our study coordinator Gesa Kix. We also thank all participating GPs and patients. Grateful thanks also to Christina Albohn-Kühne and Viktoria Bachmann for their OPTION ratings.

Funding

The study was funded by the German Ministry of Education and Research, Grant No. 01GK0701.

Conflict of interest

No conflicts of interest have been declared.

References

1 Edwards AGK, Elwyn GJ. *Evidence-Based Patient Choice – Inevitable or Impossible?* New York, NY: Oxford University Press, 2001.

2 Mazur DJ. *Shared Decision Making in the Patient-Physician Relationship: Challenges Facing Patients, Physicians, and Medical Institutions*. Tampa, FL: American College of Physician Executives, 2001.

3 Braddock CH 3rd, Edwards KA, Hasenberg NM, Laidley TL, Levinson W. Informed decision making in outpatient practice: time to get back to basics. *Journal of the American Medical Association*, 1999; **282:** 2313–2320.

4 Charles C, Gafni A, Whelan T. Shared decision-making in the medical encounter: what does it mean? (or it takes at least two to tango). *Social Science and Medicine*, 1997; **44:** 681–692.

5 Elwyn G, Edwards A, Mowle S *et al.* Measuring the involvement of patients in shared decision-making: a systematic review of instruments. *Patient Education and Counseling*, 2001; **43:** 5–22.

6 Makoul G. *Perpetuating Passivity: A Study of Physician Patient Communication and Decision Making*. Evanston, IL: Northwestern University, 1992.

7 van Thiel J, Kraan HF, Van Der Vleuten CP. Reliability and feasibility of measuring medical interviewing skills: the revised Maastricht History-Taking and Advice Checklist. *Medical Education*, 1991; **25:** 224–229.

8 Roter DL. *The Roter Method of Interaction Process Analysis*. Baltimore: The John Hopkins University, Department of Health Policy and Management, 1991.

9 Kurtz SM, Silverman JD. The Calgary-Cambridge Referenced Observation Guides: an aid to defining the curriculum and organizing the teaching in communication training programmes. *Medical Education*, 1996; **30:** 83–89.

10 Elwyn G, Edwards A, Wensing M, Hood K, Atwell C, Grol R. Shared decision making: developing the OPTION scale for measuring patient involvement. *Quality and Safety in Health Care*, 2003; **12:** 93–99.

11 Marvel MK, Schilling R, Doherty WJ, Baird MA. Levels of physician involvement with patients and their families. A model for teaching and research. *Journal of Family Practice*, 1994; **39:** 535–544.

12 Stewart M, Brown JB, Donner A *et al.* The impact of patient-centered care on outcomes. *Journal of Family Practice*, 2000; **49:** 796–804.

13 Simon D, Loh A, Harter M. Measuring (shared) decision-making–a review of psychometric instruments. *Zeitschrift für ärztliche Fortbildung und Qualitätssicherung*, 2007; **101:** 259–267.

14 Legare F, Moher D, Elwyn G, LeBlanc A, Gravel K. Instruments to assess the perception of physicians in the decision-making process of specific clinical

encounters: a systematic review. *BMC Medical Informatics and Decision Making*, 2007; **7:** 30.

15 Elwyn G, Hutchings H, Edwards A *et al.* The OPTION scale: measuring the extent that clinicians involve patients in decision-making tasks. *Health Expectations*, 2005; **8:** 34–42.

16 Goss C, Fontanesi S, Mazzi MA *et al.* Shared decision making: the reliability of the OPTION scale in Italy. *Patient Education and Counseling*, 2007; **66:** 296–302.

17 Siriwardena AN, Edwards AG, Campion P, Freeman A, Elwyn G. Involve the patient and pass the MRCGP: investigating shared decision making in a consulting skills examination using a validated instrument. *British Journal of General Practice*, 2006; **56:** 857–862.

18 Loh A, Simon D, Hennig K, Hennig B, Harter M, Elwyn G. The assessment of depressive patients' involvement in decision making in audio-taped primary care consultations. *Patient Education and Counseling*, 2006; **63:** 314–318.

19 Krones T, Keller H, Sonnichsen A *et al.* Absolute cardiovascular disease risk and shared decision making in primary care: A randomized controlled trial. *Annals of Family Medicine*, 2008; **6:** 218–227.

20 Kremer H, Ironson G. Measuring the Involvement of People with HIV in Treatment Decision Making Using the Control Preferences Scale. *Medical Decision Making*, 2008; **28:** 899–908.

21 Elwyn G, Edwards A, Kinnersley P, Grol R. Shared decision making and the concept of equipoise: the competences of involving patients in healthcare choices. *The British Journal of General Practice*, 2000; **50:** 892–899.

22 Elwyn G, Edwards A, Wensing M, Grol R. *Shared Decision Making. Measurement using the OPTION Instrument*. Wageningen: Ponsen and Looijen BV, 2005.

23 Wirtz M, Caspar F. *Beurteilerübereinstimmung und Beurteilerreliabilität.[Inter-rater Agreement and Inter-rater Reliability]*. Göttingen: Hogrefe, 2002.

24 Sprent P, Smeeton N. *Applied Nonparametric Statistical Methods*. London: Chapman and Hall, 2007.

25 Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale: Lawrence Erlbaum Associates, 1988.

26 Howell DC. *Statistical Methods for Psychology*. Florence: Cengage Learning Services, 2009.

27 Robinson BF, Bakeman R. ComKappa: A Windows 95 program for calculating kappa and related statistics. *Behavior Research Methods, Instruments, and Computers*, 1998; **30:** 731–732.

28 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*, 1977; **33:** 159–174.

29 Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized instruments in psychology. *Psychological Assessment*, 1994; **6:** 284–290.

30 Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. *Statistics in Medicine*, 1998; **17:** 101–110.

31 Bühner M. *Einführung in die Test- und Fragebogenkonstruktion [Introduction to Test and Questionnaire Construction]*. Munich: Pearson Education, 2006.

32 Pett MA, Lackey NR, Sullivan JJ. *Making Sense of Factor Analysis. The Use of Factor Analysis for Instrument Development in Health Care Research*. Thousand Oaks: Sage, 2003.

33 Swets JA. *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics: Collected Papers*. Hillsdale: Lawrence Erlbaum Associates, 1996.

34 Ellis PD. *The Essential Guide to Effect Sizes. Statistical Power, Meta-analysis, and the Interpretation of Research Results*. Cambridge: Cambridge University Press, 2010.

35 Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, 1990; **43:** 543–549.

36 Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, 1990; **43:** 551–558.

37 Weiss MC, Peters TJ. Measuring shared decision making in the consultation: a comparison of the OPTION and Informed Decision Making instruments. *Patient Education and Counseling*, 2008; **70:** 79–86.

38 Schneider A, Korner T, Mehring M, Wensing M, Elwyn G, Szecsenyi J. Impact of age, health locus of control and psychological co-morbidity on patients' preferences for shared decision making in general practice. *Patient Education and Counseling*, 2006; **61:** 292–298.

39 Melbourne E, Sinclair K, Durand MA, Legare F, Elwyn G. Developing a dyadic OPTION scale to measure perceptions of shared decision making. *Patient Education and Counseling*, 2010; **78:** 177–183.