Patient Education and Counseling 98 (2015) 970–976

Contents lists available at ScienceDirect

Patient Education and Counseling

journal homepage: www.elsevier.com/locate/pateducou

Medical Decision Making

The psychometric properties of Observer OPTION^5, an observer measure of shared decision making

Paul J. Barr [a,*], Alistair James O'Malley [a], Maka Tsulukidze [b], Michael R. Gionfriddo [c,d], Victor Montori [d], Glyn Elwyn [a,b]

[a] The Dartmouth Institute for Health Policy and Clinical Practice, Lebanon, USA
[b] The Dartmouth Center for Health Care Delivery Science, Hanover, USA
[c] Mayo Graduate School, Mayo Clinic, Rochester, USA
[d] Knowledge and Evaluation Research Unit, Mayo Clinic, Rochester, USA

ARTICLE INFO

Article history:
Received 12 December 2014
Received in revised form 15 April 2015
Accepted 18 April 2015

Keywords:
Measurement
Shared decision making
Psychometrics
Patient–provider communication

ABSTRACT

Objectives: Observer OPTION^5 was designed as a more efficient version of OPTION^{12}, the most commonly used measure of shared decision making (SDM). The current paper assesses the psychometric properties of OPTION^5.
Methods: Two raters used OPTION^5 to rate recordings of clinical encounters from two previous patient decision aid (PDA) trials ($n = 201$; $n = 110$). A subsample was re-rated two weeks later. We assessed discriminative validity, inter-rater reliability, intra-rater reliability, and concurrent validity.
Results: OPTION^5 demonstrated discriminative validity, with increases in SDM between usual care and PDA arms. OPTION^5 also demonstrated concurrent validity with OPTION^{12}, $r = 0.61$ (95%CI 0.54, 0.68) and intra-rater reliability, $r = 0.93$ (0.83, 0.97). The mean difference in rater score was 8.89 (95% Credibility Interval, 7.5, 10.3), with intraclass correlation (ICC) of 0.67 (95% Credibility Interval, 0.51, 0.91) for the accuracy of rater scores and 0.70 (95% Credibility Interval, 0.56, 0.94) for the consistency of rater scores across encounters, indicating good inter-rater reliability. Raters reported lower cognitive burden when using OPTION^5 compared to OPTION^{12}.
Conclusions: OPTION^5 is a brief, theoretically grounded observer measure of SDM with promising psychometric properties in this sample and low burden on raters.
Practice implications: OPTION^5 has potential to provide reliable, valid assessment of SDM in clinical encounters.

1. Introduction

1.1. Background

Brief patient reported measures are the most scalable means of assessing the level of shared decision making (SDM) in routine clinical encounters. However, patient reports can be subject to halo, leniency and gratitude biases, which tend to provide unvarying high scores (ceiling effects) that make it difficult to discriminate between high and low achievement of SDM [1]. Observer (direct) measures, where trained assessors directly observe behavior via, for example, recorded (audio, video, transcript) encounters, are theoretically less prone to these effects, and can be triangulated with patient reported measures to produce more comprehensive assessment [2].

There are several observer measures of SDM, ranging 7–70 items [3–9], with variable levels of reliability and largely unexamined validity [10]. Observer OPTION^{12} (OPTION^{12}) [6] is the most commonly-used observer measure. *OPTION^{12} is a one-dimensional scale, consisting of 12 items, which focuses specifically on clinician behavior*. While discriminative validity of OPTION^{12} is satisfactory, inter-rater reliability *has been mixed* [11]. *In addition, OPTION^{12} has also been criticized for it lack of focus on the elicitation of patient preference*. Some specified behaviors are rarely observed

Abbreviations: DSAT-10, Decision Support Analysis Tool; FRAX, Fracture Risk Assessment Tool; ICC, intraclass correlation coefficient; PDA, patient decision aid; SDM, shared decision-making; SD, standard deviation.

* Corresponding author at: 35 Centerra Dr, Lebanon, NH 03766, USA. Tel.: +1 603 646 2578; fax: +1 603 646 2268.
E-mail addresses: Paul.J.Barr@dartmouth.edu (P.J. Barr), Alistair.J.O'Malley@dartmouth.edu (A.J. O'Malley), tsulukidze.maka@gmail.com (M. Tsulukidze), Gionfriddo.Michael@mayo.edu (M.R. Gionfriddo), Montori.Victor@mayo.edu (V. Montori), glynelwyn@gmail.com (G. Elwyn).

http://dx.doi.org/10.1016/j.pec.2015.04.010
0738-3991/

(Item 2, Item 3 and Item 10). *In addition, assessor burden is high due to the high number of items, which may not occur in a linear fashion during a medical encounter*. In hindsight, Elwyn et al. [12] believe OPTION[12] contained some items depicting an idealized form of SDM that is unrealistic in the real world, and missed items that address other core aspects, such as dealing with patient preferences [13].

Observer OPTION[5] was developed as a five-item measure to address issues with OPTION[12] (http://www.optioninstrument.org/) with fewer items and a focus on the assessment of patient preferences [12]. OPTION[5] is based on Collaborative Deliberation, a *conceptual* model *describing the process of patients considering alternative health care options, in collaboration with clinicians and others. While existing theories of health care communication focus on decision making of a single individual, the model of Collaborative Deliberation was an attempt to develop a model that considers a collaborative effort. The model* consists of five core dimensions: (1) constructive interpersonal engagement, (2) recognition of alternative actions, (3) comparative learning, (4) preference construction and elicitation, and (5) preference integration [14]. The goal was to provide a more efficient measure, that focused more on the core components of SDM while retaining good psychometric properties.

1.2. Aims, objectives and hypotheses

This study aimed to assess the psychometric qualities of OPTION[5] using video and audio recorded clinical encounters from two trials, the Osteoporosis Choice Randomized Trials [15] and the Chest Pain Choice Trial [16].

2. Methods

2.1. Study design

Data was analyzed from two trials assessing the impact of patient decision aids (PDAs) used during the medical encounter: the Osteoporosis Choice Randomized trial [15] and the Chest Pain Choice trial [16]. In the Osteoporosis Choice trial, postmenopausal women in primary care aged 50 years or older, at risk for osteoporotic fractures, and eligible for bisphosphonate therapy were randomized to an intervention or usual care arm. The World Health Organization's Fracture Risk Assessment Tool (FRAX®) was introduced partway through the trial, providing patients in the intervention arm with a fracture risk assessment. Accounting for this change, we grouped the patients from Osteoporosis Choice into two groups, Osteoporosis I (no FRAX®) and Osteoporosis II (FRAX®). In the Chest Pain Choice trial, adults with chest pain who were being considered for admission for prolonged observation and cardiac stress test in a specialized unit were randomly allocated to receive a patient decision aid or usual care. For summaries of each trial, see Table 1. The study was approved by the Mayo Clinic Institutional Review Board (#13-004057).

2.2. Observer measures of SDM

2.2.1. Description of OPTION[12]

OPTION[12] was used to assess the level of SDM in the original two trials. OPTION[12] is a twelve-item scale with each item rated from '0' to '4', where '0' represents absence of an SDM-specific competency and '4' represents optimal performance [6]. Item scores are added (maximum of 48) and re-scaled to a value between 0 and 100. A minimum of two raters is suggested; raters perform independent assessments of audio- or video-recorded encounters, and mean scores are calculated.

2.2.2. Description of OPTION[5]

The five items of OPTION[5] replicate the response format, assessment and scoring methods of OPTION[12] (Table 2; Appendix 1).

2.2.3. Standardized training and calibration of raters in the use of Observer OPTION[5]

Two raters from the Knowledge and Evaluation Research Unit, Mayo Clinic, previously trained in OPTION[12], assessed the recorded

Table 1
Characteristics of participants and OPTION[12] score of the encounters.[a]

	Chest pain		Osteoporosis I		Osteoporosis II	
	Usual Care	PDA	Usual care	PDA	Usual care	PDA
N	100	101	35	37	25	13
Video	90 (90%)	92 (91%)	31 (89%)	34 (92%)	19 (76%)	12 (92%)
Audio	10 (10%)	9 (9%)	4 (11%)	3 (8%)	6 (24%)	1 (8%)
Part of encounter recorded[b]	N/A	N/A				
Whole encounter			17 (48%)	21 (57%)	6 (24%)	1 (8%)
Discussion only			2 (6%)	5 (13%)	11 (44%)	11 (84%)
Through discussion			16 (46%)	11 (30%)	8 (32%)	1 (8%)
Video length						
Mean (SD), Median (Range)	3.4 (2.3), 3 (1, 12)	5.3 (2.4), 5 (1, 11)	21.2 (15), 17 (2, 59)	21.2 (13), 20 (3, 54)	15.4 (15), 11 (2, 55)	11.2 (4), 11 (5, 21)
Female: *N* (%)	59 (59%)	59 (58%)	35 (100%)	37 (100%)	25 (100%)	13 (100%)
Caucasian: *N* (%)	97 (97%)	92 (91%)	35 (100%)	37 (100%)	25 (100%)	12 (92%)
Age						
Mean (SD), (Range)	54.6 (12), (29, 82)	54.5 (12), (27, 87)	65.7 (9), (50, 82)	66.4 (9), (51, 84)	63.5 (11), (50, 86)	70.2 (6), (55, 79)
Education						
≤Highschool	24 (24%)	27 (27%)	10 (28.6%)	12 (32.4%)	6 (24%)	4 (31%)
>Highschool	75 (75%)	71 (70%)	25 (71.4%)	24 (64.9%)	18 (62%)	8 (61%)
Missing	1 (1%)	3 (3%)	0 (–)	1 (2.7%)	1 (4%)	1 (8%)
OPTION[12] Mean score (0–100)	7	26.6	26.9	49.9	42.6	57.4
(95% CI)	(6, 8)	(25, 28)	(22, 32)	(43, 57)	(37, 48)	(50, 64)

PDA: decision aid; UC: usual care.
[a] All statistics presented (mean, SD and CI) are unadjusted.
[b] Indicates whether the video records the whole encounter, just the discussion of interest or the whole encounter up until the end of the discussion of interest. This coding was not conducted for the Chest Pain trial.

Table 2
Items included in OPTION[5].

Item number – dimension	Item content
Item 1 – Justify the work of deliberation	The provider draws attention to, or re-affirms, a problem where alternate treatment or management options exist and which requires the initiation of a decision making process. If the patient draws attention to the availability of options and the provider responds by agreeing that the options need consideration, the item can also be scored positively
Item 2 – Justify the work of supporting deliberation	The provider reassures the patient, or re-affirms, that the provider will support the patient to become informed. The provider supports/explains the need to deliberate about the options
Item 3 – Inform, describe options, exchange views	The provider gives information, or re-affirms/checks understanding, about options that are considered reasonable (including taking 'no action'), to support the patient in understanding/comparing the pros and the cons
Item 4 – Elicit preferences	The provider supports the patient to examine, voice, and explore his/her personal preferences in response to the options that have been described
Item 5 – Integrate preferences	The provider makes an effort to integrate the patient's preferences as decisions are either made by the patient or arrived at by a process of collaboration/discussion

encounters using Observer OPTION[5]. MT supervised the training process. The raters were given: (1) the OPTION[5] development paper [12] and user manual and (2) five encounter recordings (taken from both intervention and usual care arms) to independently assess. Their initial scores were discussed, and a second set of five encounters assessed and the scores compared. Disagreements on scores were resolved based on discussion with *a member of the research team (MT).*

2.3. Data collection and management

After training, all encounters from the trials were independently assessed using OPTION[5]. To assess intra-rater reliability, a random subsample of encounters was re-rated within two weeks of initial assessment, *with a target of 30 per rater*. OPTION[5] ratings were added to prior trial data (OPTION[12] scores, clinician characteristics and patient gender and education level) and stored anonymously in a REDCap system [17].

2.4. Psychometric data analysis

Statistical analyses are outlined in Table 3; the primary goal was to assess the discriminative validity of OPTION[5]. OPTION[12] scores in the trials were significantly higher where PDAs were used, so we hypothesized that OPTION[5] scores would act similarly and show high discriminative validity. Mean OPTION[5] scores from trial arms were compared using an independent group Student's *T*-Test analysis. Further analyses assessed concurrent validity, intra-rater reliability and inter-rater reliability (Table 3).

To estimate the inter-rater intraclass correlation coefficient (ICC) for the OPTION[5] scores of the two raters, we used WinBUGS statistical software [21] to fit a three-level hierarchical model with random effects for encounter and study, covariates for trial arm and rater, and a quadratic variance function to account for differences in the variance (heteroscedasticity) when the OPTION[5] score is close to 0 compared to the scale mid-point of 50. Because heteroscedasticity makes inter-rater reliability dependent on the score level, we averaged inter-rater ICC calculations across the three trials. We calculated both the ICC accuracy of raters' score agreement and ICC consistency of raters' scores (their ability to achieve the same rank order) [22]. We calculated both the ICC accuracy of raters' score agreement and ICC consistency of raters' scores (their ability to achieve the same rank order) and fitted 95% credibility intervals [23].

2.5. Power calculation

No prior data existed for OPTION[5], so we used the standard deviation from OPTION[12] studies [6], specifically SD 7.68, to estimate the minimum difference detectable between trial arms. We estimated that the sample size available for analysis would provide 90% power to detect, each with an alpha level of 0.05, (1) a minimum of a 3.5 point difference in OPTION[5] scores in the Chest Pain trial, (2) a minimum of a 6 point difference in OPTION[5] scores in Osteoporosis I, and (3) a minimum of a 10 point difference in OPTION[5] scores in Osteoporosis II.

3. Results

Across the randomized trials, 151 patients received PDA interventions and 160 received usual care. Participant characteristics are outlined in Table 1.

Table 3
Psychometric analyses of OPTION[5].[a]

Psychometric property	Definition	Assessment	Analyses
Discriminative validity (Construct validity)	Ability to yield low scores when the construct under measurement is absent, and higher scores as the presence of the construct increases [1]	Between-group comparisons	Between groups *T*-test or Welch test
Concurrent validity [1] (Criterion validity)	Presence of correlation between measures that claim to assess the same construct [1]	Relationship between OPTION[5] and OPTION[12] scores	Pearson product moment correlation (*r*) [18]
Inter-rater reliability (Overall score)	The degree of concordance between raters' scores on the same encounter	Relationship between raters' OPTION[5] scores	Intraclass correlation coefficients, accounting for known differences between encounters (trial arm, the amount of SDM, study effects)
Inter-rater reliability (Item by item)	The degree of concordance between raters' item scores on the same the encounter	Relationship between scores for individual items of OPTION[5]	Weighted Cohen's Kappa coefficients [19]
Intra-rater reliability	Consistency of ratings, of the same encounter, across two time points by the same rater [1]	Relationship between scores at two time points	Pearson product moment correlation (*r*) [18]

[a] Based on the criteria outlined by Jarvis et al. [20], OPTION[5] is considered a formative measure, where a change in one item does not necessarily lead to a change in other items. Therefore an assessment of internal consistency is not required.

Table 4
OPTION[5] overall score and breakdown by items.

	Chest pain		Osteoporosis I		Osteoporosis II	
	Usual care	Patient decision aid	Usual care	Patient decision aid	Usual care	Patient decision aid
OPTION[5] – *overall score*						
Mean (SD)	9.78 (12.8)	44.97 (10.82)	27.5 (14.5)	39.39 (12.3)	26.5 (15.9)	43.27 (8.4)
Median (Range)	2.5 (0–52.5)	42.5 (17.5–72.5)	25 (10–72.5)	40 (5–60)	27.5 (2.5–55)	40 (35–60)
T Statistic	−20.72		−3.65		−4.24[a]	
p-Value	<0.001		<0.001		<0.001	
OPTION[5] – *Breakdown by items*						
Item 1						
Mean (SD)	0.42 (0.68)	2.16 (.40)	1.44 (0.54)	1.84 (0.51)	1.54 (0.69)	1.96 (0.52)
Median (Range)	0 (0–3)	2 (0–3)	2 (1–3)	2 (1–3)	2 (1–3)	2 (1–3)
Item 2						
Mean (SD)	0.14 (0.32)	1.01 (0.60)	0.65 (0.59)	1 (0.5)	0.44 (0.33)	0.77 (0.44)
Median (Range)	0 (0–2)	1 (0–3)	1 (0–3)	1 (0–2)	1 (0–1)	1 (1–2)
Item 3						
Mean (SD)	0.34 (0.55)	1.95 (0.59)	1.22 (0.71)	1.91 (0.70)	1.3 (0.74)	2.27 (0.33)
Median (Range)	0 (0–2)	2 (0–4)	1 (0–3)	2 (0–3)	2 (0–3)	2 (2–3)
Item 4						
Mean (SD)	0.47 (0.66)	1.86 (0.56)	1.11 (0.77)	1.62 (0.62)	1 (0.88)	1.85 (0.55)
Median (Range)	0 (0–3)	2 (1–4)	1 (0–3)	2 (0–3)	1 (0–3)	2 (1–3)
Item 5						
Mean (SD)	0.6 (0.64)	1.99 (0.53)	0.98 (0.77)	1.51 (0.63)	1.02 (0.87)	1.81 (0.48)
Median (Range)	1 (0–3)	2 (1–3)	1 (0–3)	2 (0–3)	1 (0–3)	2 (1–3)

[a] Reports results of Welch test due to unequal variances between arms.

3.1. Discriminative validity

OPTION[5] demonstrated discriminative validity. The OPTION[5] mean rating for patients receiving the Chest Pain Choice decision aid was 35.2 points higher than that of the usual care group ($p < 0.001$). Similarly, statistically significant differences were observed in OPTION[5] scores between trial arms in the Osteoporosis I and II groups, with the two PDA arms' OPTION[5] scores 11.9 points and 16.7 points higher, respectively ($p < 0.001$) (Table 4).

Higher scores were achieved on OPTION[5] in the PDA arm of the Chest Pain Choice trial [16] compared to the original OPTION[12] scores for this arm. Yet, the opposite occurred in the Osteoporosis trial [15], with the original OPTION[12] scores higher than those reported using OPTION[5]. In addition, scores in the usual care arm of Osteoporosis II were considerably higher with the original OPTION[12] compared to OPTION[5].

3.2. Item performance

SDM behaviors (Table 4) were rarely observed in the Chest Pain trial's usual care arm, with a median score of 0 (no effort) for all but one item. In the PDA group, the median score increased to 2 (baseline effort) for all but one item. The Osteoporosis groups had a higher level of SDM in usual care, with a median score of 1 (minimal effort) commonly recorded. In all trials, item scores in the PDA group were higher than in the usual care arm.

3.3. Concurrent validity, intra-rater and inter-rater reliability

The concurrent validity of OPTION[5] was demonstrated by its moderate positive correlation with OPTION[12], $r = 0.61$ (95% Credibility Interval 0.54, 0.68). Intra-rater reliability was also demonstrated with a strong positive correlation, $r = 0.93$ (95% Credibility Interval 0.83, 0.97), between time 1 and time 2 ratings of the same encounter; however, only one rater completed this task for a total of 22 encounters. The inter-rater reliability at the item level, across all trials, using a weighted Cohen's Kappa statistic ranged from fair (items 2 and 5) to substantial (item 3) agreement (Table 5).

3.3.1. Overall inter-rater reliability

While rater 1 tended to give higher ratings to individual clinical encounters than rater 2, Fig. 1 highlights the consistency of their rankings from least to most SDM. Rater agreement was highest when OPTION[5] scores were closer to 0. As OPTION[5] scores within an encounter approached the midpoint, 50, the level of agreement between raters was at its lowest (Appendix 2). In order to estimate the ICC to account for the variation in rater agreement, we constructed a hierarchical model consisting of three levels: (i) Trial (Chest Pain Choice, Osteoporosis I and II); (ii) Use of PDA; and (iii) Rater. We added a quadratic variance function to account for clustering and heteroscedasticity (see Section 2). The mean difference in the raters' OPTION[5] scores was estimated to be 8.89 (95% Credibility Interval, 7.48 to 10.30). The ICC for the accuracy of raters' scores across encounters was 0.66 (95% Credibility Interval, 0.51, 0.91) and the ICC for the consistency of rater scores across encounters was 0.70 (95% Credibility Interval, 0.56 to 0.94).

Table 5
Weighted Cohen's Kappa by item.

Item	Agreement	Expected Agreement	Kappa	Agreement	95% CI	*p*-Value
Item 1	87.0%	71.4%	0.54	Moderate	0.46, 0.60	<0.0001
Item 2	82.3%	77.6%	0.21	Fair	0.16, 0.25	<0.0001
Item 3	90.1%	72.6%	0.64	Substantial	0.60, 0.67	<0.0001
Item 4	86.0%	72.7%	0.49	Moderate	0.44, 0.50	<0.0001
Item 5	75.5%	60.9%	0.37	Fair	0.35, 0.40	<0.0001

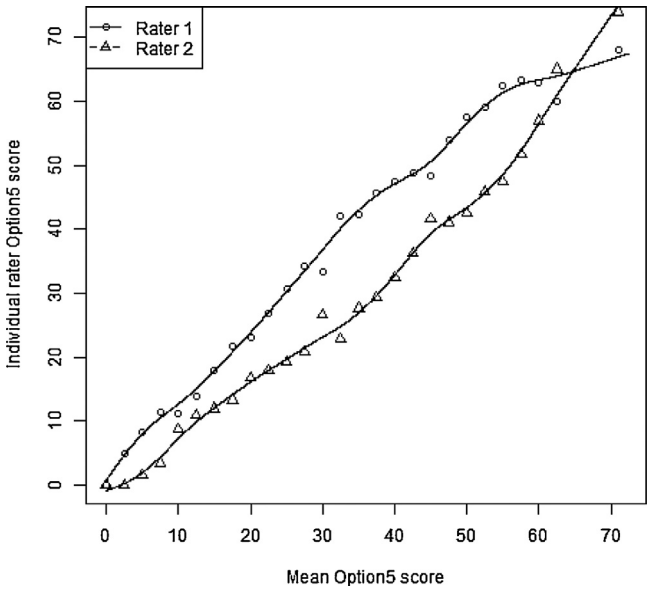


Fig. 1. Individual rater scores tracked against mean OPTION[5] score by encounter (across trials).

3.4. Rater feedback

The raters reported that OPTION[5] imposed less cognitive burden than OPTION[12] due to the brevity of the scale and the improved differentiation of the target behaviors. Raters suggested revisions to the manual, such as providing examples of expressions eligible for a positive score. Raters indicated that item 2 was difficult to score and suggested refinement of its wording. At the end of the study, the *items* were revised to allow patient-led or clinician-led behaviors to be scored positively (Appendix 3).

4. Discussion and conclusion

4.1. Discussion

In a sample of audio and video data from clinical encounters in two randomized trials, OPTION[5] produced valid and reliable estimates of SDM. OPTION[5] demonstrated discriminative validity, concurrent validity with OPTION[12], intra-rater reliability and promising inter-rater reliability at the item level. Raters welcomed the improved differentiation between target behaviors and the reduced cognitive burden of the tool compared to the 12-item version.

We acknowledge several study limitations. New raters, i.e., not those who performed the original OPTION[12] assessments, completed the OPTION[5] assessments. However, we estimate that these independent ratings are likely to have underestimated the relationship between the two measures. In addition, the use of an additional observer measure would provide even more support for concurrent validity, but this was not feasible in the current study *due to lack of resources*. Secondly, the rating process occurred over a period of 8 months; however, there was no opportunity to reassess the raters' approach to applying OPTION[5]. Rater application of OPTION[5] may have changed over this time period, affecting the consistency of their ratings. Future projects should periodically evaluate rater performance and offer refresher-training sessions. Finally, we did not have data about the encounter duration in the trials at the time of analysis. *The rating of items may not occur in a sequential order as medical encounters are dynamic and vary from patient to patient. Encounters of longer duration may be more challenging to rate as they require raters to maintain high levels of concentration for longer periods*. We hypothesize that encounters of greater duration may lead to reduced inter-rater agreement. *To protect against this hypothesized effect raters should be trained extensively with exposure to many types of encounter and rating sessions should be limited to a set time that maximizes rater concentration levels*. This requires further investigation.

At the individual item level, items demonstrated higher observed SDM performance in the intervention arms of the trials. Item 2 (justifying the work of deliberation as a team) was the only item that remained minimally observed in the intervention arms, and had the lowest levels of inter-rater agreement. Item 2 was reported to be the most unclear, an issue that required modification in a revised OPTION[5] scale. Similar items have been reported problematic in other observer measures, for example, the item "Intervening to help patients handle pressure or inadequate support from others" contained in the Decision Support Analysis Tool (DSAT-10) [9]. Improving the item specificity and providing more guidance may help improve inter-rater agreement. However, if the behavior is inherently rarely observed, it is likely to have a low Kappa score despite high agreement between raters [24]. Given the categorical nature of the judgments, we cannot assume equal intervals between response points but nevertheless, detailed guidance of how to score at each response level would lead to improved inter-rater reliability.

The overall inter-rater reliability of OPTION[5] was comparable to other observer measures of SDM. As recommended with OPTION[12], assessment with two independent raters is suggested for research purposes. Using two raters may reduce the influence of lenient or harsh assessments, the so called "dove and hawk" effect, although statistical adjustments can be made if this phenomenon is observed [25].

OPTION[5] demonstrated a moderate positive correlation with OPTION[12]. OPTION[5] scores were much higher than the original OPTION[12] scores in the PDA group for the Chest Pain trial, yet the opposite was observed with Osteoporosis trial groups where OPTION[12] scores were higher than those recorded using OPTION[5]. As the Chest Pain trial included short encounters (3–5 min), encounter duration may have precluded clinicians from exhibiting all 12 behaviors listed in OPTION[12]. A positive relationship between total encounter duration and higher scores on OPTION[12] has been previously identified.[11] It is possible that the sole focus on core elements of SDM in OPTION[5] attenuates this relationship; further research is required. The lower scores observed using OPTION[5] in the Osteoporosis trial may reflect items from OPTION[12] that were scored highly not being included in OPTION[5]. For example, item 9 from OPTION[12] (offering opportunities for questions) is one of the scale's most frequently-observed behaviors [11].

Finally, the inter-rater reliability of OPTION[12] in the original trials as measured using the Lin correlation coefficient [26] was over 0.9. We opted to pool the estimate for OPTION[5] and account for heteroscedasticity, allowing more generalizable estimates of inter-rater reliability, and found lower, but still adequate, inter-rater reliability (ICC ranging from 0.66 to 0.70). The assessment of inter-rater reliability is often done without checking the validity of assumptions associated with such tests; for example, constant error variance is assumed when estimating inter-rater reliability using ICCs. Our pooled estimates alleviate concerns related to violation of standard test assumptions.

4.2. Practice implications

OPTION[5] is the shortest existing observer measure of SDM. This study shows that it maintains comparable psychometric

performance to other related measures [10]. An assessment of measure validity and direct comparison between observer measures has not been frequently undertaken in this field [10]. In this project, we demonstrated the concurrent and discriminative validity of OPTION^5 using encounter data from patients participating in two trials, indicating high construct and criterion validity. This is crucial, as it is possible to achieve high reliability without validity, but you may not be measuring the construct of true interest [1]. The version of OPTION^5 used in this study primarily assesses clinicians' behavior, although item 1 allows for consideration of a patient's contribution to option recognition. Allowing positive scoring when the patient leads the participation process is vital provided the clinician is supportive; this needs to be addressed in a revised OPTION^5 scale (Appendix 3) [10]. Another unique feature of OPTION^5 is the explicit assessment of preference integration, arguably the crux of SDM (item 5).

OPTION^5 presents an opportunity to provide feedback to clinicians on how well they involve patients in decisions, and could assist with training efforts. Due to its relative efficiency, OPTION^5 could increase the feasibility of assessing clinical encounters, possibly using real-time observations. In routine practice, patient reported measures could act as screening tools for SDM, coupled with observer measures, such as OPTION^5, to provide more detailed information on practice where SDM levels are low. This would also allow for investigation of correlations between observed and perceived assessments of SDM, a current gap in research.

4.3. Conclusion

OPTION^5 is a brief, theoretically grounded observer measure of SDM that has the potential to provide reliable, valid assessment of SDM in clinical encounters. Further testing of a refined OPTION^5 scale is required.

Conflicts of interest

No external support was received for this work. Glyn Elwyn has received funding from the Informed Medical Decisions Foundation, Boston, MA, USA, and provides ad hoc consulting to Emmi Solutions, Chicago, IL, USA. MRG was supported by CTSA Grant Number TL1 TR000137 from the National Center for Advancing Translational Science (NCATS). This manuscript's contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH. No other authors have conflicts of interests to report. Prof. Elwyn and Dr. Barr wish to declare this intellectual conflict of interest. CollaboRATE is also freely available under a Creative Commons License for non-commercial use – CC BY-NC-ND 3.0 Unported. CollaboRATE is available under license for commercial organizations, to date no fees have been levied for this.

Acknowledgments

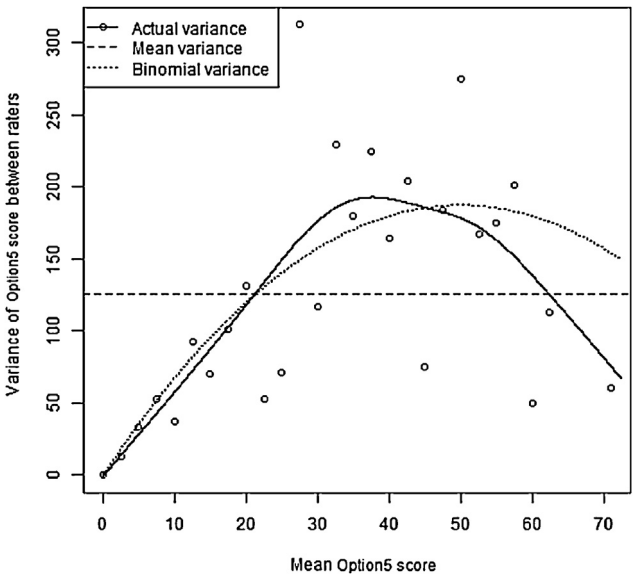
We would like to acknowledge Dr. Annie Le Blanc and Dr. Megan Branda for their comments on the study design, and Angela Sivly for her work as an OPTION^5 rater.

Appendix 1. Original version of OPTION^5

The following items form Observer OPTION^5. Items should be scored independently of each other. Scoring should be summed so the potential total score is between 0 and 20. We advocate rescaling this score to lie between 0 and 100.

Observer OPTION^5 Measure	Score
Item 1: The provider *draws attention to, or re-affirms*, a problem where alternate treatment or management options exist, and which requires the initiation of a decision making process. If the patient draws attention to the availability of options, and the provider responds by agreeing that the options need consideration, the item can also be scored positively. 0 = No effort 1 = Minimal effort 2 = Baseline effort 3 = Skilled effort 4 = Exemplary effort	
Item 2: The provider reassures the patient, or re-affirms, that the provider will support the patient *to become informed*. The provider will support/explain the *need to* deliberate about the options. 0 = No effort 1 = Minimal effort 2 = Baseline effort 3 = Skilled effort 4 = Exemplary effort	
Item 3: The provider gives information, or re-affirms/checks understanding, about options that are considered reasonable (including taking 'no action'), to support the patient in understanding/comparing the pros and cons. 0 = No effort 1 = Minimal effort 2 = Baseline effort 3 = Skilled effort 4 = Exemplary effort	
Item 4: The provider supports the patient to examine, voice, and explore his/her personal preferences in response to the options that have been described. 0 = No effort 1 = Minimal effort 2 = Baseline effort 3 = Skilled effort 4 = Exemplary effort	
Item 5: The provider makes an *effort to integrate* the patient's preferences as decisions are either made by the patient or arrived at by a process of collaboration and discussion 0 = No effort 1 = Minimal effort 2 = Baseline effort 3 = Skilled effort 4 = Exemplary effort	
Total Score 0–20 Rescale 0–100	

Appendix 2. Variance of OPTION[5] scores between raters plotted against mean OPTION[5] score by encounter (across trials)



Appendix 3. Modified version; changes are italicized

The following items form Observer OPTION 5. Items should be scored independently of each other. Scoring should be summed so the total score is between 0 and 20, *and then rescaled to lie between 0 and 100.*

Observer OPTION [5] Measure	Score

Item 1: *For the health issue being discussed, the clinician* draws attention to or re-affirms *that* alternate treatment or management options exist or *that the need for a decision exists*. If the patient *rather than the clinician* draws attention to the availability of options, *the clinician* responds by agreeing that the options need *deliberation.*
0 = No effort 1 = Minimal effort *2 = Moderate effort*
3 = Skilled effort 4 = Exemplary effort

Item 2: *The clinician* reassures the patient, or re-affirms, that *the clinician* will support the patient to become informed *and to deliberate about the options. If the patient states that they have sought or obtained information prior to the encounter, the clinician supports such a deliberation process.*
0 = No effort 1 = Minimal effort 2 = *Moderate effort*
3 = Skilled effort 4 = Exemplary effort

Item 3: *The clinician* gives information, or checks understanding, about *the pros and cons* of the options that are considered reasonable (including taking 'no action'), *to support the patient in comparing the alternatives. If the patient requests clarification, explores options, or compares options, the clinician supports the process.*

0 = No effort 1 = Minimal effort 2 = *Moderate effort*
3 = Skilled effort 4 = Exemplary effort

Item 4: *The clinician makes an effort to elicit the patient's preferences* in response to the options that have been described. *If the patient declares their preference(s), the clinician is receptive/supportive.*
0 = No effort 1 = Minimal effort 2 = *Moderate effort*
3 = Skilled effort 4 = Exemplary effort

Item 5: *The clinician* makes an effort to integrate the patient's preferences as decisions are made. *If the patient indicates how best to integrate their preferences as decisions are made, the clinician is supportive.*
0 = No effort 1 = Minimal effort 2 = *Moderate effort*
3 = Skilled effort 4 = Exemplary effort

Total Score 0–20
Rescale 0–100

References

[1] Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. New York: Oxford University Press; 2003.
[2] McCroskey J, McCroskey L. Self report as an approach to measuring communication competence. Commun Res Rep 1988;5:113.
[3] Braddock III CH, Edwards KA, Hasenberg NM, Laidley TL, Levinson W. Informed decision making in outpatient practice. J Am Med Assoc 1999;282:2313.
[4] Brown RF, Butow PN, Juraskova I, Ribi K, Gerber D, Bernhard J, et al. Sharing decisions in breast cancer care: development of the Decision Analysis System for Oncology (DAS-O) to identify shared decision making during treatment consultations. Health Expect 2011;14:29–37.
[5] Butow P, Juraskova I, Chang S, Lopez AL, Brown R, Bernhard J. Shared decision making coding systems: how do they compare in the oncology context? Patient Educ Couns 2010;78:261–8.
[6] Elwyn G, Hutchings H, Edwards A, Rapport F, Wensing M, Cheung WY, et al. The OPTION scale: measuring the extent that clinicians involve patients in decision-making tasks. Health Expect 2005;8:34–42.
[7] Guimond P, Bunn H, O'Connor AM, Jacobsen MJ, Tait VK, Drake ER, et al. Validation of a tool to assess health practitioners' decision support and communication skills. Patient Educ Couns 2003;50:235–45.
[8] Singh S, Butow P, Charles M, Tattersall MHN. Shared decision making in oncology: assessing oncologist behaviour in consultations in which adjuvant therapy is considered after primary surgical treatment. Health Expect 2010;13:244–57.
[9] Stacey D, Taljaard M, Drake ER, O'Connor AM. Audit and feedback using the brief Decision Support Analysis Tool (DSAT-10) to evaluate nurse-standardized patient encounters. Patient Educ Couns 2008;73:519–25.
[10] Scholl I, Koelewijn-van Loon M, Sepucha K, Elwyn G, Légaré F, Härter M, et al. Measurement of shared decision making: a review of instruments. Z Evid Fortbild Qual Gesundhwes 2011;105:313–24.
[11] Couët N, Desroches S, Robitaille H, Vaillancourt H, Leblanc A, Turcotte S, et al. Assessments of the extent to which health-care providers involve patients in decision making: a systematic review of studies using the OPTION instrument. Health Expect 2013. http://dx.doi.org/10.1111/hex.12054.
[12] Elwyn G, Tsulukidze M, Edwards A, Légaré F, Newcombe R. Using a talk model of shared decision making to propose an observation-based measure: observer OPTION 5 Item. Patient Educ Couns 2013;93:265–71.
[13] Glasgow RE, Riley WT. Pragmatic measures: what they are and why we need them. Am J Prev Med 2013;45:237–43.
[14] Elwyn G, Lloyd A, May C, van der Weijden T, Stiggelbout A, Edwards A, et al. Collaborative deliberation: a model for patient care. Patient Educ Couns 2014;97:158–64.
[15] Montori VM, Shah ND, Pencille LJ, Branda ME, Van Houten HK, Swiglo BA, et al. Use of a decision aid to improve treatment decisions in osteoporosis: the osteoporosis choice randomized trial. Am J Med 2011;124:549–56.
[16] Hess EP, Knoedler MA, Shah ND, Kline JA, Breslin M, Branda ME, et al. The chest pain choice decision aid: a randomized trial. Circ Cardiovasc Qual Outcomes 2012;5:251–9.
[17] Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap) – a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009;42:377–81.
[18] Cohen J. A power primer. Quant Methods Psychol 1992;112:155–9.
[19] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.
[20] Jarvis C, MacKenzie S, Podsakoff P. A critical review of construct indicators and measurement model misspecification in marketing and consumer research. J Consum Res 2003;30:199–218.
[21] Spiegelhalter D, Thomas A, Best N. WinBUGS; 1999.
[22] Carrasco J, Jover J. The concordance correlation coefficient estimated through variance components. In: IX Conf. Española Biometría. 2003. p. 1–4.
[23] Eberly LE, Casella G. Estimating Bayesian credible intervals. J Stat Plan Inference 2003;112:115–32.
[24] Feinstein AR, Cicchetti DV. High agreement but low Kappa: I. The problems of two paradoxes. J Clin Epidemiol 1990;43:543–9.
[25] Harasym PH, Woloschuk W, Cunning L. Undesired variance due to examiner stringency/leniency effect in communication skill scores assessed in OSCEs. Adv Health Sci Educ Theory Pract 2008;13:617–32.
[26] Steichen T, Cox N. A note on the concordance correlation coefficient. Stata J 2002;2:183–9.