OPEN ACCESS

Development and validation of the Overall Fidelity Enactment Scale for Complex Interventions (OFES-CI)

Liane Ginsburg,[1] Matthias Hoben,[2] Whitney Berta,[3] Malcolm Doupe,[4,5] Carole A Estabrooks,[2] Peter G Norton,[6] Colin Reid,[7] Ariane Geerts,[8] Adrian Wagg[9]

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjqs-2023-016001).

For numbered affiliations see end of article.

Correspondence to
Dr Liane Ginsburg, Health Policy & Management, York University Faculty of Health, Toronto, ON M3J 1P3, Canada; lgins@yorku.ca

Received 1 February 2023
Accepted 5 August 2023

To cite: Ginsburg L, Hoben M, Berta W, *et al*. *BMJ Qual Saf* Epub ahead of print: [please include Day Month Year]. doi:10.1136/bmjqs-2023-016001

ABSTRACT

Background In many quality improvement (QI) and other complex interventions, assessing the fidelity with which participants 'enact' intervention activities (ie, implement them as intended) is underexplored. Adapting the evaluative approach used in objective structured clinical examinations, we aimed to develop and validate a practical approach to assessing fidelity enactment—the Overall Fidelity Enactment Scale for Complex Interventions (OFES-CI).

Methods We developed the OFES-CI to evaluate enactment of the SCOPE QI intervention, which teaches nursing home teams to use plan-do-study-act (PDSA) cycles. The OFES-CI was piloted and revised early in SCOPE with good inter-rater reliability, so we proceeded with a single rater. An intraclass correlation coefficient (ICC) was used to assess inter-rater reliability. For 27 SCOPE teams, we used ICC to compare two methods for assessing fidelity enactment: (1) OFES-CI ratings provided by one of five trained experts who observed structured 6 min PDSA progress presentations made at the end of SCOPE, (2) average rating of two coders' deductive content analysis of qualitative process evaluation data collected during the final 3 months of SCOPE (our gold standard).

Results Using Cicchetti's classification, inter-rater reliability between two coders who derived the gold standard enactment score was 'excellent' (ICC=0.93, 95% CI=0.85 to 0.97). Inter-rater reliability between the OFES-CI and the gold standard was good (ICC=0.71, 95% CI=0.46 to 0.86), after removing one team where open-text comments were discrepant with the rating. Rater feedback suggests the OFES-CI has strong face validity and positive implementation qualities (acceptability, easy to use, low training requirements).

Conclusions The OFES-CI provides a promising novel approach for assessing fidelity enactment in QI and other complex interventions. It demonstrates good reliability against our gold standard assessment approach and addresses the practicality problem in fidelity assessment by virtue of its suitable implementation qualities. Steps for adapting the OFES-CI to other complex interventions are offered.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ There is a growing knowledge base regarding how to assess the fidelity with which quality improvement (QI) and other complex interventions are delivered, though there is relatively little knowledge regarding how to efficiently assess the fidelity with which they are implemented (enacted) by intervention participants. Data on fidelity enactment is critical for proper interpretation of intervention outcomes.

WHAT THIS STUDY ADDS

⇒ The present study developed and validated an easy-to-use, robust approach for assessment of fidelity enactment for use in QI and other complex interventions (the Overall Fidelity Enactment Scale for Complex Interventions (OFES-CI)) and outlines specific procedures for assessing fidelity.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ The OFES-CI can be easily adapted for practical assessment of fidelity of other complex interventions. Such fidelity data can help address well-known problems with intervention replication by providing valuable insight into why interventions succeed or fail and what adaptations may be needed to promote greater success.

BACKGROUND

When an evaluation shows that an intervention or quality improvement (QI) initiative did not achieve its aims, it is often hard to know if this means the intervention is ineffective or it was simply not implemented as planned. Fidelity of a QI or other intervention reflects the extent to which that intervention is implemented as intended[1] and its assessment is extremely important. Ignoring fidelity increases the risk of discarding potentially effective interventions that failed to work because they were not properly implemented or accepting ineffective interventions whose outcomes were brought about by factors other than the intervention.[2 3]

With some interventions/QI initiatives, assessing fidelity is straightforward. For example, in a trial in which an order set is implemented to improve care for patients with diabetes, one could assess fidelity simply by looking at how often the order set was used for eligible patients. Assessing fidelity is not so straightforward with more complex interventions[4 5] and QI programmes, such as testing the use of team-based plan-do-study-act (PDSA) cycles to improve care for nursing home residents. With complex interventions and QI programmes (such as use of PDSA cycles where proper implementation is known to be challenging[6–9]), there are often multiple interacting components, multiple actors, and fidelity often involves implementing a series of ongoing activities. In these instances, it is useful to consider fidelity frameworks,[10 11] which differentiate between fidelity delivery (ie, consistent delivery, as per protocol, to target persons who are to implement behaviours of interest), fidelity receipt (intervention participants' comprehension of intervention behaviours and capacity to use the skills taught) and fidelity enactment which is the focus of the current study and reflects actual performance of intervention skills/implementation of the core components of an intervention or QI programme.

With more complex interventions, audio or video recording and coding is generally recognised to be the gold standard for assessing fidelity delivery.[12] However, expert assessment of recorded activities is costly and, more importantly, it is largely infeasible for assessing fidelity enactment in complex/pragmatic interventions since it is impractical for researchers to record or observe teams on an ongoing basis as they enact intervention skills/activities.[13] With complex interventions, fidelity enactment is sometimes assessed using audit, observation or detailed self-report checklists containing items that reflect core components of the intervention. However, each of these approaches carries its own challenges pertaining to cost and/or bias.

Fidelity enactment of QI and other complex interventions is underexplored.[4 10 14 15] The need for efficient,[16] high-quality, practical approaches to assessment of fidelity enactment has been highlighted by several recent reviews,[4 12 15 17] as has the need for studies that outline specific procedures for assessing fidelity.[15 18 19] Building on our previous work,[20–22] this study aimed to: (1) develop an easy to use, objective approach to the assessment of fidelity enactment—the Overall Fidelity Enactment Scale for use in Complex Interventions (the OFES-CI)—and, (2) validate it by comparing its results with gold standard fidelity enactment scores gleaned from detailed process evaluation data. Our development and validation work was carried out in the context of assessing teams' ability to carry out PDSA approaches to improve resident care in nursing homes during the SCOPE QI intervention study[23] (see box 1 for a description and schematic summarising SCOPE).

The proposed approach to assessing fidelity enactment is an adaptation of the evaluative approach used in objective structured clinical examinations (OSCEs). OSCEs are routinely used to assess competency of health professional trainees prior to entry to practice. In an OSCE, trainees interact with standardised patients in a series of 5–10 min encounters during which the trainee must demonstrate competency by assessing or resolving a clinical problem. These encounters are observed and evaluated by clinicians who rate the level of competency that the trainee demonstrates during the encounter. In the proposed approach, rather than rating trainees as they interact with standardised patients, subject matter experts rated teams' presentations of PDSA progress in the SCOPE intervention. The proposed approach is supported by the OSCE literature, which has shown that (a) subject matter experts are able to reliably evaluate holistic skills in the context of a brief interaction,[24 25] and (b) global assessment scales may have higher reliability and may be more sensitive to variation in intervention participant skills than assessing discrete skills on a checklist.[24 26] Our approach is also supported by psychology and counselling research which suggests that assessing fidelity (usually delivery of complex treatment regimens) becomes more difficult as an intervention becomes less prescriptive and expert raters, given their experience, can appropriately use discretion to accept minor variations on intervention fidelity.[27]

METHODS

Design

We developed an overall measure of fidelity enactment (the OFES-CI) and then validated it using secondary data collected as part of a process evaluation of the SCOPE intervention.[28] Specifically (and described in detail below), we compared the OFES-CI ratings obtained from experts who observed PDSA progress presentations made at the end of SCOPE to more detailed and comprehensive qualitative process evaluation data collected during the final 3 months of the intervention (our gold standard).

Setting—the SCOPE intervention

The SCOPE intervention (summarised in box 1) is a complex intervention conducted in 31 nursing homes from four health regions in Western Canada in 2018-2019 which aimed to achieve quality improvement using the breakthrough series model.[29] SCOPE is delivered primarily by a QI lead and teaches teams, led by healthcare aides, to enact/implement PDSA

Box 1 The SCOPE intervention with schematic

⇒ SCOPE is modelled on the Institute for Healthcare Improvement's Breakthrough Series Collaborative Model[29] and was designed to be implementable. Using the PARiHS framework,[42 43] SCOPE addresses technical aspects of conducting a PDSA cycle, provides facilitation and addresses contextual factors necessary to support implementation.

⇒ SCOPE trial outcomes included best practice use and improvement in the clinical area that teams chose to work on: pain, responsive behaviours or mobility. Outcomes were measured using Resident Assessment Instrument–Minimum Data Set (RAI-MDS 2.0) indicators.[44]

⇒ The year-long intervention began in June 2018 in four health regions in the Canadian provinces of Alberta and British Columbia. Each of the 31 nursing homes had one unit-based improvement team. Teams had five to seven members, were led by a healthcare aide and included at least two healthcare aides.

⇒ Teams attended quarterly learning congresses (LCs) with other teams in their region to network and participate in plenary sessions and activities on the improvement model, measurement in PDSA cycles and team dynamics. Teams presented on project progress at the second, third and fourth LCs.

⇒ Teams received support from a team sponsor (unit manager) and a senior sponsor (nursing home director). Teams received coaching from a quality advisor (QA) to support quality improvement (QI) activities and instil a new approach to improvement work at the bedside. Researchers in geriatrics, nursing, implementation science, QI and health services supported the quality team.

⇒ A mixed-methods concurrent process evaluation was conducted.[28] Process data collected and intervals are shown on the bottom of the schematic below.

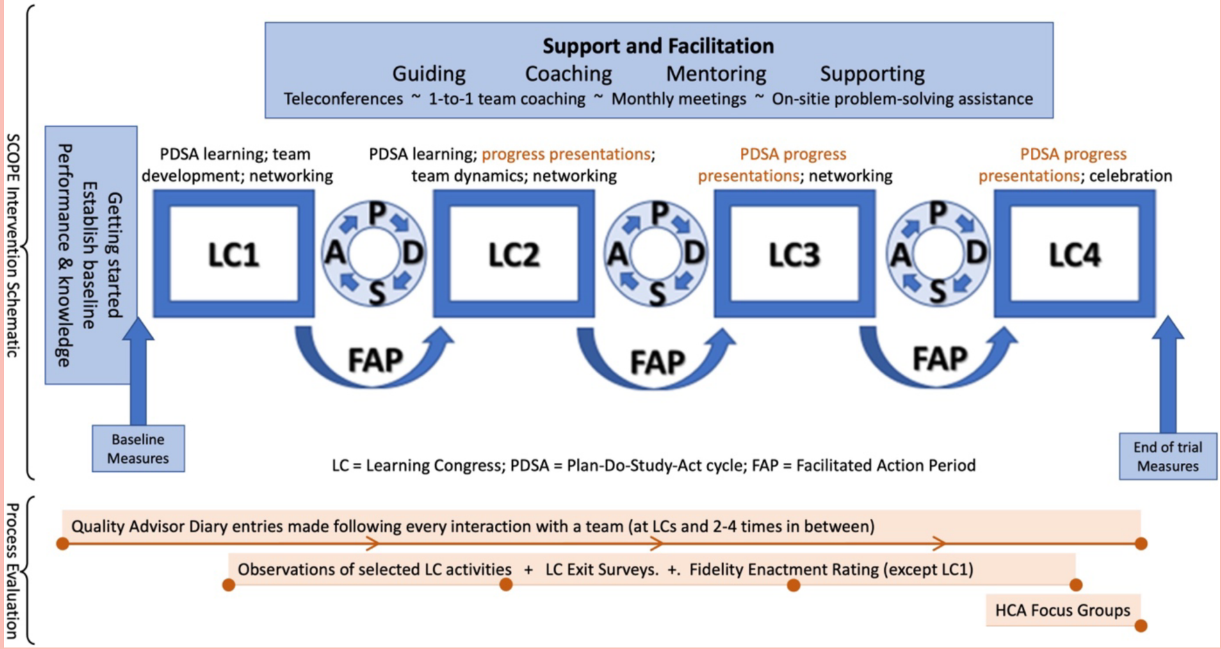
⇒ The core components of the intervention include:

⇒ SCOPE is a multicomponent pragmatic trial at the level of the resident care team in 31 nursing homes. SCOPE teaches local Healthcare Aide-led teams to implement improvement initiatives based on current best evidence.[23] SCOPE is unique in engaging and equipping healthcare aides to lead an improvement team.

1. Care aide-led teams working on a focused clinical area
2. Use of quality improvement methods by unit teams (change concepts, measurement, Plan-Do-Study-Act cycles)

The current study assessed enactment of these components

3. In-person meetings with all teams (quarterly learning congresses (LCs))
4. Ongoing support from a Quality Advisor during action periods between LCs
5. Supporting leaders to facilitate and support change and the care aide-led teams


SCOPE Intervention Schematic
Support and Facilitation
Guiding
Coaching
Mentoring
Supporting
Teleconferences ~ 1-to-1 team coaching ~ Monthly meetings ~ On-sitie problem-solving assistance
Getting started
Establish baseline
Performance & knowledge
PDSA learning; team development; networking
PDSA learning; progress presentations; team dynamics; networking
PDSA progress presentations; networking
PDSA progress presentations; celebration
LC1
LC2
LC3
LC4
P
D
S
A
FAP
Baseline Measures
End of trial Measures
LC = Learning Congress; PDSA = Plan-Do-Study-Act cycle; FAP = Facilitated Action Period
Process Evaluation
Quality Advisor Diary entries made following every interaction with a team (at LCs and 2-4 times in between)
Observations of selected LC activities + LC Exit Surveys. +. Fidelity Enactment Rating (except LC1)
HCA Focus Groups

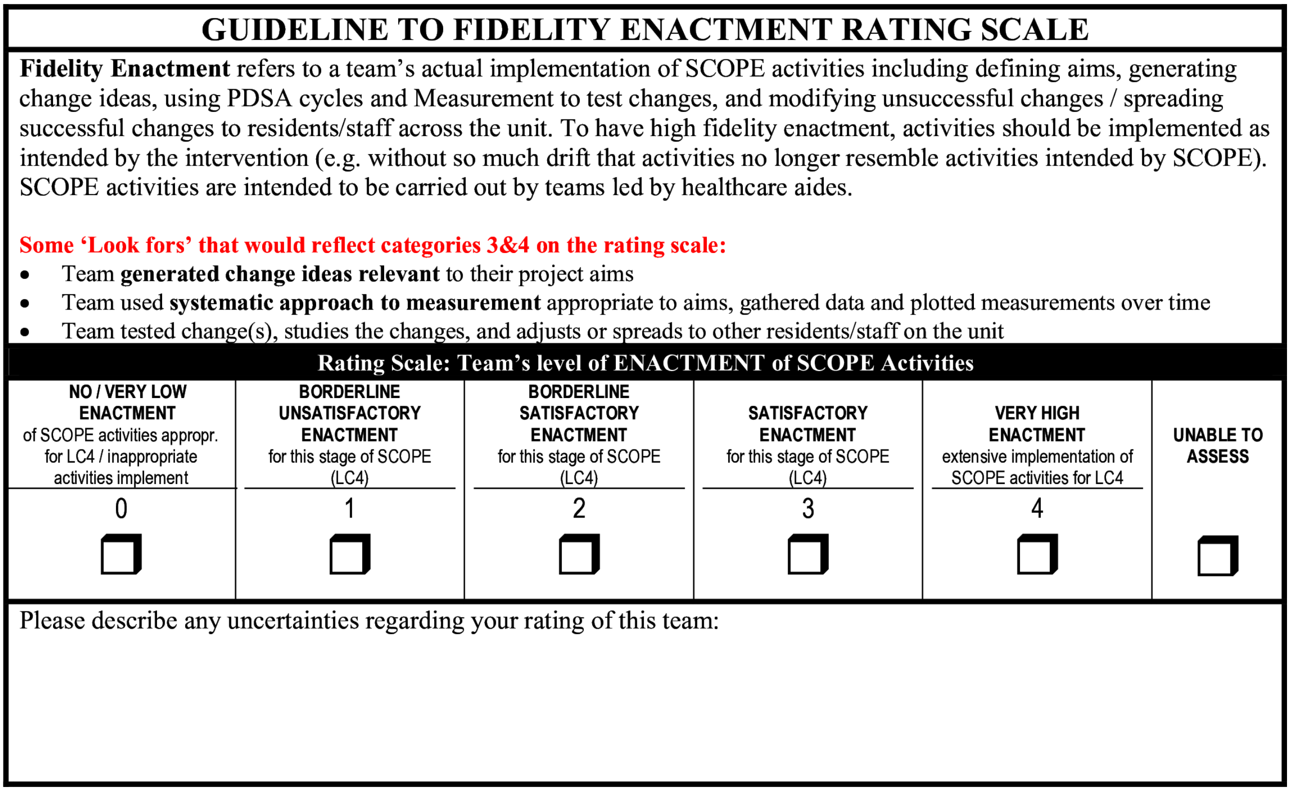
cycles to improve resident care. During the 1-year intervention, teams participated in quarterly learning congresses (LCs) conducted in each region where the PDSA approach was taught (LC1) and reinforced (LC2). Healthcare aid-led teams were expected to implement PDSA cycles between LCs with internal facilitation from local facility leaders and QI-specific facilitation from an external quality advisor (QA). Teams presented their PDSA implementation progress at LCs 2–4. All SCOPE activities and LCs took place in-person. The SCOPE trial[23] and process evaluation[28] are published elsewhere.

Development of the OFES-CI

The OFES-CI was developed alongside SCOPE following steps outlined by Walton and colleagues[12] for developing high-quality fidelity measures. We also adhered to practices used in our previous work on fidelity assessment[21 30] and on the use of expert raters.[22] As a first step, the core components of the SCOPE intervention (see box 1) were analysed by the first two authors to specify activities that were intended to be enacted by each healthcare aide-led team. These core components and activities included: (1) use of a unit-based team, led by healthcare aides, to work on one of three clinical areas (pain, mobility, behaviour) and, (2) use of specific QI methods taught during SCOPE related to aim development, change concepts, measurement and PDSA cycles. Next, we drafted a single-item overall measure of fidelity enactment, the OFES-CI, that incorporated the components and activities in (1) and (2). In keeping with the OSCE assessment approach, 'Guidelines for rating' that include a definition of what constitutes fidelity enactment in SCOPE and 'look fors' that reflect activities appropriate for the upper two categories on the rating scale were included in the OFES-CI. The OFES-CI was used to assess the level of fidelity enactment at the second, third and fourth (final) Learning Congresses (LC). It uses a 5-point rating scale where a rating of '0' indicates 'No/Very low enactment of scope activities appropriate for [LC#]/inappropriate activities implemented' and a rating of '4' indicates 'Very high enactment—extensive implementation of SCOPE activities for [LC#]'. The OFES-CI was developed in the first quarter of SCOPE. We obtained feedback from SCOPE researchers about its content, wording and face validity, and pilot tested the approach at the second LC (see below). Figure 1 shows the OFES-CI used at the final LC (LC4).

Data collection requirements using the OFES-CI

For experts to rate fidelity using the OFES-CI, we had to provide opportunities in SCOPE for teams to demonstrate the extent and ways in which they had implemented the core intervention components. As noted, the SCOPE intervention included four, quarterly, LCs and at the second, third, and fourth congresses each team gave a structured 6 min 'progress presentation' where they were asked specifically to describe (a) what improvement activities they had undertaken during the previous quarter, including details of the PDSA cycles they conducted, and (b) what data they collected to know whether their efforts were leading to improvement. We treated these LC progress presentations as



GUIDELINE TO FIDELITY ENACTMENT RATING SCALE

Fidelity Enactment refers to a team's actual implementation of SCOPE activities including defining aims, generating change ideas, using PDSA cycles and Measurement to test changes, and modifying unsuccessful changes / spreading successful changes to residents/staff across the unit. To have high fidelity enactment, activities should be implemented as intended by the intervention (e.g. without so much drift that activities no longer resemble activities intended by SCOPE). SCOPE activities are intended to be carried out by teams led by healthcare aides.

Some 'Look fors' that would reflect categories 3&4 on the rating scale:

- Team **generated change ideas relevant** to their project aims
- Team used **systematic approach to measurement** appropriate to aims, gathered data and plotted measurements over time
- Team tested change(s), studies the changes, and adjusts or spreads to other residents/staff on the unit

Rating Scale: Team's level of ENACTMENT of SCOPE Activities

NO / VERY LOW ENACTMENT of SCOPE activities appropr. for LC4 / inappropriate activities implement	BORDERLINE UNSATISFACTORY ENACTMENT for this stage of SCOPE (LC4)	BORDERLINE SATISFACTORY ENACTMENT for this stage of SCOPE (LC4)	SATISFACTORY ENACTMENT for this stage of SCOPE (LC4)	VERY HIGH ENACTMENT extensive implementation of SCOPE activities for LC4	UNABLE TO ASSESS
0	1	2	3	4	
□	□	□	□	□	□

Please describe any uncertainties regarding your rating of this team:

Figure 1 The OFES-CI global fidelity enactment measure. This fidelity rating scale was applied to project presentations teams gave at learning congresses (LCs) 2–4. The full OFES-CI package for raters (with instructions and the actual form) can be found in the online supplemental appendix. OFES-CI, Overall Fidelity Enactment Scale for Complex Interventions; PDSA, plan-do-study-act.

analogous to an OSCE standardised patient encounter and applied a similar evaluative approach—an expert rater observed the 6 min progress presentation, asked clarification questions, and then completed the OFES-CI based on their observations.

Expert raters were members of the SCOPE investigator team from different provinces with expertise in geriatrics, implementation science and/or improvement science. They were all familiar with SCOPE, QI and the concept of fidelity enactment. Raters attended LCs on the date(s)/in the region(s) most convenient for them, so the same expert did not rate all teams. For global measures like the OFES-CI, we followed guidelines from OSCE research regarding the need for raters to (a) have clear instructions and evaluation criteria and (b) be sufficiently trained and calibrated.[25]

Pilot testing the OFES-CI and rater training

We pilot tested the OFES-CI with all 31 SCOPE teams at the second LC (LC2) in each health region by having two experts provide an enactment rating for each team's LC2 PDSA progress presentation. Prior to LC2, all raters conducted pre-work and participated in a 30 min zoom training session led by the first author. To ensure raters had a common understanding of the 'Guidelines for rating', the training session reviewed the definition of fidelity enactment in SCOPE, the rating scale categories and the 'look fors', and it included a calibration exercise. Inter-rater reliability of the two experts' LC2 OFES-CI ratings was assessed using a one-way random effects consistency intraclass correlation coefficient (ICC) (appropriate when the same pair of raters is not used for all teams) and was found to be good[31] (ICC=0.73, 95% CI=0.43 to 0.87). Based on this result we proceeded with a single expert rater at the third and fourth LCs.

We used the OFES-CI with all teams at the third LC. The rater debrief yielded feedback regarding OFES-CI acceptability and usability and also suggested four additional ways to improve the OFES-CI that were incorporated into the LC4 rating process: (1) we added a short Q&A following each progress presentation where raters were encouraged to ask a question to better enable them to assess fidelity enactment; (2) since some raters were overly strict in their LC3 assessment of measurement in a PDSA cycle, we conducted an additional training session prior to LC4 and included calibration scenarios for discussion; (3) a 0.5 rating (between two categories) was added so that raters did not feel overly constrained by the 5-point rating scale. They were also asked if they might 'raise/lower their rating by ½ or 1 category'; (4) a comment box was added so raters could qualify or explain any ratings they were unsure about. All LC2 and LC4 rater training materials, as well as the final OFES-CI package with rater instructions, are included as online supplemental material for interested readers.

Sample

Twenty-seven of 31 SCOPE teams attended the final LC (LC4). An OFES-CI rating was collected for each of these 27 teams. Ratings were provided by one of five experts who were trained in the manner described above. Each expert provided ratings for 3–7 teams (raters who attended LC4 in one region rated 3–4 teams; raters who attended LC4 in two regions rated 6–7 teams).

Validating the OFES-CI—procedures and analysis

Arriving at our 'gold standard'

Coding of detailed qualitative process evaluation data is an approach which has been used previously to assess PDSA cycle fidelity[9] and may be the closest we can get to a gold standard approach to assessing fidelity enactment. Throughout SCOPE, team-specific process evaluation data were collected to facilitate understanding of the extent and ways in which teams implemented the intervention (see bottom of box 1 schematic). To arrive at a 'gold standard' fidelity enactment rating for the current study, we made use of the following process evaluation data[28] collected between the end of the third and fourth LCs: (1) QA diary entries made each time the QA was in contact with a team, (2) responses to open-ended questions provided by SCOPE participants on LC exit surveys, (3) observations conducted by trained members of the research team of various LC activities. Table 1 provides details about these three sources of data, which amounted to several pages of rich textual data for each team between LCs 3 and 4. We arrived at our 'gold standard' fidelity enactment rating in the fall of 2021 using the following three steps:

Step 1. We conducted a calibration exercise using process evaluation data for three teams, collected during the 3-month period leading up to the third LC. The aim was to see whether three authors (LG, WB, MH) could independently code the qualitative data using deductive content analysis[32] against the OFES-CI categories and arrive at consensus. Comparisons between coders led to minor scale clarification discussions.

Step 2. The same three authors independently coded qualitative data for five teams, this time for the 3-month period leading up to the final LC. The aim, for coders to achieve ratings that were within 1 point of each other on the 5-point OFES-CI scale, was achieved for 4/5 teams. Coders differing by 1.5 points for the fifth team. Inter-rater reliability was examined using a two-way mixed consistency average measures ICC, appropriate for estimating the reliability of the mean ratings provided by the same set of coders for ordinal data.[33] The ICC was excellent for these five teams (0.95, 95% CI=0.84 to 0.99), enabling us to proceed to step 3.[31]

Step 3. Again using deductive content analysis, the remaining 22 teams were coded by two of the

Table 1 SCOPE process evaluation data used to arrive at gold standard fidelity enactment rating

	Data collection approach	Purpose	Data used to arrive at the gold standard
QA Diaries	QAs completed a diary entry after each interaction with a team. Diary entries included both facts (eg, 'Team tested X change'; 'x# of team members attended') and impressions (eg, 'physio seems to be driving changes for this team').	Diaries were intended to capture QAs' perspectives regarding team engagement, progress, challenges, enactment of SCOPE's core components, deviations from intended practice (adaptations) and the role of context in implementation.	2–4 diary entries per team between LC3 and LC4, including one entry immediately after LC4. Entries following a QI support session or following the LC were typically ~500 words.
Learning Congress Exit Surveys	Exit surveys completed at each learning congress by care aides, leaders and QAs included open-ended questions that were included with each team's qualitative process data.	Open-ended items sought team-level data from multiple stakeholder perspectives regarding SCOPE acceptability, extent implementation was care aide led, team accomplishments, support received, and implementation facilitators and barriers.	78 surveys from SCOPE team members (avg. 2.9 completed surveys per team at LC4); surveys from 39 leaders (25 unit managers and 14 directors of care at LC4), one QA survey for each team they supported in SCOPE.
Observations	Researchers were trained to use a semi-structured template to observe teams in certain LC activities. Activities captured pertinent processes (leadership approaches, team dynamics) and fidelity receipt/enactment of QI methods central to SCOPE.	To capture qualitative data, primarily about the processes through which teams engage and interact with the intervention.	Two researchers observed two different activities for each team at LC4: presentation of team storyboards and PDSA progress presentation.

LC, learning congress; PDSA, plan-do-study-act; QA, quality advisor; QI, quality improvement.

authors—LG and either WB or MH. Both coders independently applied the OFES-CI categories to the qualitative data for the 3-month period leading up to the final LC then discussed any cases where ratings were more than 1 point apart. Coders were always blinded to team names. Inter-rater reliability between the two coders for all 27 teams that participated in the final LC (5 teams coded in step 2 and 22 teams coded in step 3) was examined using a one-way random effects average measures ICC, appropriate since teams were not all coded by the same pair of coders.[33] For each team, coders' scores were averaged to create a 'gold standard' enactment rating. The gold standard therefore reflects an enactment rating based on review of detailed qualitative data on SCOPE implementation activities that took place during the final 3 months of the intervention.

Validating the OFES-CI against the 'gold standard'

We validated the OFES-CI ratings collected from the 27 teams at the final LC (Spring 2019) against the gold standard. A one-way random effects single measures ICC (appropriate when not all pairs of ratings are provided by the same coders[34]) was used to compare the expert OFES-CI rating of the PDSA progress presentation with the gold standard enactment rating derived using steps 1–3 above. This single measures ICC provides a measure of reliability of the OFES-CI when used by one subject matter expert in the context of a time-limited interaction at the end of an intervention. For interpretation of all ICCs, we used the classification proposed by Cicchetti[31] (inter-rater reliability less than 0.40 is poor; 0.40–0.59 is fair; 0.60–0.74 is good; 0.75–1.00 is excellent).

RESULTS

OFES-CI implementation qualities

Informal feedback from SCOPE researchers on the initial draft of the OFES-CI and from the pilot indicated the tool appeared to represent the construct it is supposed to be measuring (SCOPE fidelity enactment), suggesting strong face validity. Feedback from the pilot and the LC3 rater debrief clearly indicated acceptability—all raters noted the tool is quick to use (low burden) and easy to apply to PDSA progress presentations, particularly if comments and ratings between categories are permitted.

Generating the gold standard fidelity enactment rating

Inter-rater reliability (step 3 above) was excellent (one-way random effects average measures ICC=0.93, 95% CI=0.85 to 0.97), indicating that coders had high agreement in their application of the OFES-CI categories to the qualitative data. We therefore used the average score provided by two coders as the gold standard fidelity enactment rating for each team.

Validating the OFES-CI against the gold standard fidelity enactment rating

Inter-rater reliability, performed to assess the degree to which the OFES-CI expert rating was consistent with the gold standard enactment rating, was 'fair' (one-way random effects single measures ICC=0.58, 95% CI=0.26 to 0.78). There was one team with a gold standard enactment rating of 0.25 ('No/Very low enactment of SCOPE activities') and an OFES-CI expert rating of 4.0 ('Very High Enactment'). A comment on the OFES-CI rating form for this team stated that 'They are doing gigantic amounts of stuff…but they seem to have done so before SCOPE … I REALLY wonder to what extent we can attribute the good ratings above

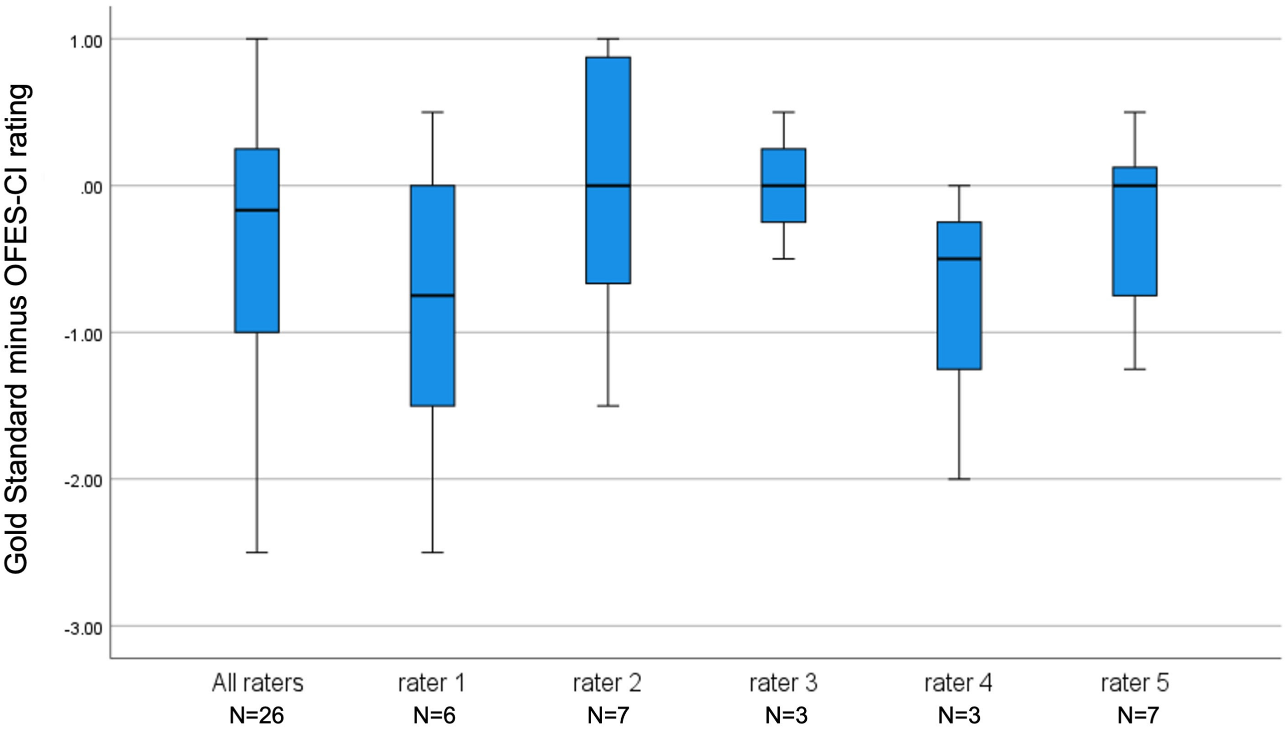


Figure 2 Gold standard fidelity enactment rating and OFES-CI rating difference scores. OFES-CI, Overall Fidelity Enactment Scale for Complex Interventions.

[the OFES-CI ratings] to SCOPE… [several initiatives described] …were already successful - how much has SCOPE added????'. Unfortunately, these comments were not reviewed immediately following the final LC (in which case we would have reminded the rater that their rating should reflect activities enacted as part of SCOPE and invited them to revise it). Because this was an error in the research process rather than the OFES-CI rating process, we removed this case from our analysis (final n=26). After removing data from this team, inter-rater reliability was 'good' (ICC=0.71, 95% CI=0.46 to 0.86).

Nine of the final 26 OFES-CI ratings included certainty adjustments (recall raters could indicate they might raise or lower their rating by 0.5 or 1 category). We examined their effects by adjusting the OFES-CI rating up or down by half a point for these nine cases. The ICC remained unchanged when these adjustments were included (ICC=0.70, 95% CI=0.440.85).

As a final analysis, we looked for evidence of any systematic differences between the OFES-CI rating and the gold standard rating (ie, was the gold standard always higher or lower?) and between the five raters. The OFES-CI ratings (mean=2.62, SD=1.3, range 0.0–4.0) and the gold standard fidelity enactment ratings (mean=2.25, SD=1.2, range 0.5–4.0) both reflect use of the full 0–4 rating scale for the final 26 cases. Figure 2 shows the distribution of gold standard and OFES-CI rating difference scores for all 26 cases (far left boxplot) and for each rater. The mean difference between the two ratings is −0.37 (median difference=−0.17) indicating the gold standard ratings were, on average, 0.37 points lower than the OFES-CI ratings. The left boxplot also shows that 75% of the gold standard and OFES-CI ratings were within 1 point of each other. None of the individual expert's OFES-CI ratings were systematically higher or lower than the gold standard rating.

DISCUSSION

Fidelity enactment is an important indicator of implementation success.[16] Its assessment can provide considerable insight regarding the potential value of QI and other complex initiatives/interventions. This study builds on robust approaches to assessment used in medical education[25] and describes the development and validation of the OFES-CI. The OFES-CI offers a sound and judicious approach to assessing fidelity enactment that is not currently found in the literature. The approach demonstrates good reliability against our gold standard assessment after removal of one case where the open text was not consistent with the OFES rating given. The OFES-CI addresses the practicality problem in fidelity assessment[30] by virtue of its suitable implementation qualities (acceptability, ease of completion, low burden, low training requirements).

Similar to Walton's findings,[12] our piloting, training and calibration work support the importance of these processes in the development and application of any fidelity enactment measure. Pre-testing the OFES-CI during the second and third LCs suggested useful refinements to the tool and the data collection process—of these, we suggest retaining the comments box and allowing ratings between categories to enhance usability. Our findings indicate the certainty adjustment added after LC3 is probably not required. Piloting and training, including the use of calibration activities, may be particularly important for global fidelity enactment measures like the OFES-CI that assess the enactment of multiple intervention components in a single measure. We also concur with Walton's suggestion that clear definitions of what constitutes fidelity enactment must be provided to expert raters to limit individual judgement and subjectivity.[12]

Our validation analysis comparing the OFES-CI ratings to the gold standard (objective 2) identified one large discrepancy, described above, where the OFES-CI rating indicated very high enactment while the gold standard rating suggested no or very low enactment. Researchers using the OFES-CI approach are strongly encouraged to include the open-text field to permit raters to qualify their ratings if necessary. Importantly, OFES-CI rating forms should be checked by a member of the research team immediately following completion to identify any instances where qualitative comments do not match the rating provided, so that discrepancies can be resolved. Our failure to review the qualitative comments resulted in a missing OFES-CI rating for one of the teams in our analysis. Studies of complex group-level or organization-level QI interventions, even large ones, often do not have large samples[35] and its therefore crucial to minimise missing data.[30]

The need for validated fidelity enactment tools and practical guidance for their use was identified by 70–80% of researchers surveyed in a recent study.[36] The OFES-CI approach can meet the needs of researchers and those testing QI interventions by overcoming three practical and methodological challenges associated with assessment of fidelity enactment: (1) the absence of a gold standard approach for measuring fidelity receipt or enactment[12] (though we contend that collecting and coding detailed process evaluation data may offer one such approach); (2) fidelity enactment, as typically assessed using participant self-report checklists, has unclear reliability and validity and low concordance with observer ratings[17]; (3) fidelity measures, including enactment measures, need to be specific to intervention skills and their measurement properties are therefore rarely established.[12]

Practice implications

The OFES-CI can be helpful for those involved in QI. We can be more confident about a QI initiative that appears to be effective if we also have high OFES-CI scores, indicating the initiative was implemented with fidelity. Similarly, when a QI initiative appears not to have the intended effects, OFES-CI scores can help sort out whether it is an effectiveness problem or an implementation problem—that is, high OFES-CI scores suggest an effectiveness problem, lower OFES-CI scores suggest implementation challenges that may (or may not) be readily overcome. Even greater insights may accrue if the OFES-CI is used along with other process evaluation data and/or if raters are asked to use the OFES-CI open text box to comment on which intervention components or activities participants struggled with. Pinpointing components participants struggled to implement can suggest what adaptations may be required to improve the intervention, its implementation and/or its scale up.

The OFES-CI development process we describe is generalisable—it can be adapted to assess enactment of a variety of complex QI and other interventions. Box 2 outlines steps for creating an OFES-CI that is specific to other study contexts. Importantly, these steps should be undertaken concurrently with the development of the intervention. In addition, all steps will be accomplished best by individuals with intimate knowledge of the intervention or QI initiative whose fidelity is being assessed, provided due consideration is given to the benefits and potential biases associated with using the same researchers in the design of an interventions, its evaluation and the evaluation of fidelity (see Moore *et al*[3] for an important discussion of these trade-offs). Lastly, step III requires some flexibility in the structure of an intervention (so opportunities for participants to demonstrate fidelity enactment can be built in). When evaluating the fidelity of initiatives that are replications of established interventions, it will be important to ensure processes introduced to facilitate fidelity assessment do not substantively alter the intervention under study.[30]

While the OFES-CI would benefit from further validation in other intervention contexts, we suggest that study teams can use the OFES-CI approach to understand and quantify fidelity enactment in QI and other complex interventions without undertaking the validation procedures and analysis we conducted using the gold standard. Indeed, previous work by this team using the OFES-CI approach, without the validation work described here, showed evidence of its predictive validity in the INFORM trial where overall fidelity enactment was positively associated with improvements in the primary study outcome (formal team communications).[21] Ultimately, the OFES-CI approach, on its own or conducted as part of a larger process evaluation, can strengthen the analysis and interpretation of QI and other intervention data.[3]

Those adapting the OFES-CI should be aware that the fidelity measurement process is not easy and may even amount to a small parallel study.[30] Although we contend that the OFES-CI is reliable and relatively easy to use to rate fidelity enactment, the adaptation process outlined in box 2 must be carried out thoughtfully and may require additional measurement or fidelity expertise. This is particularly true for step III, when opportunities to demonstrate fidelity enactment are built

Box 2 Steps for adapting the OFES-CI to other study contexts

Step I. Identify primary intervention target participant(s) whose enactment activities will be assessed (in SCOPE it was unit teams led by healthcare aids).
Step II. Identify core components of the intervention (ie, skills and/or activities) to be enactment by participants (from step I) to achieve fidelity to the intervention. Include these in a definition of fidelity enactment for the new intervention that will, ultimately, be included on the OFES-CI form. List approximately 2–4 things that expert raters would 'look for' as evidence of successful enactment.
Step III. Outline potential ways to build opportunities to assess fidelity enactment into the intervention. In SCOPE we used progress presentations. Other approaches could involve asking intervention participants to deliver a short teaching session (or make a video) demonstrating how they might teach intervention skills/how to implement intervention activities to their peers. Like in an OSCE, these should ideally be brief instances built into the intervention where participants can demonstrate that they have acquired intervention skills and/or enacted key intervention activities from step II. This step may require the most attention and creativity in the OFES adaptation process.
Step IV. Adapt the OFES-CI form presented here to the new context using the fidelity definition and 'look fors' generated in step II. Maintain the 5-point rating scale categories ('very low/no enactment' to 'very high enactment') commonly used in OSCEs; retain the comment box so raters can qualify ratings if need be. Allow the use of ratings between two categories to enhance usability.
Step V. Solicit feedback, train expert raters and pilot the adapted OFES-CI to promote clarity regarding what constitutes high–low fidelity in the new intervention/QI context. The feedback, training and piloting should ideally (a) enable discussion of what constitutes fidelity (eg, what activities or skills and at what level of proficiency), and (b) include a calibration activity (eg, using a mock case or video if there are no natural calibration opportunities early in the intervention). Rater training information, fidelity definitions and calibration approaches included in the online supplemental material can be used as a guide to these step V activities.

OFES-CI, Overall Fidelity Enactment Scale for Complex Interventions; OSCEs, objective structured clinical examinations; QI, quality improvement.

into the intervention. If step III is not done thoughtfully, the adapted OFES-CI may have low sensitivity (eg, true enactment of an initiative may be high but said enactment may not be evident from the presentation/other opportunity created to demonstrate enactment) or low specificity (eg, true enactment is low, but participants are able to exaggerate their efforts). When opportunities to demonstrate fidelity enactment are built into the intervention, it is important they are as structured as possible to improve both sensitivity and specificity of the OFES-CI.

Study limitations and future research

Studies of inter-rater reliability should be designed so that ratings are independent to avoid inflating ICCs.[37] In the current validation study, the three authors who coded the qualitative data to come up with the gold standard also provided some of the OFES-CI ratings at the final LC. Potential for bias is mitigated by two factors[37]—coders were blind to team names when coding the qualitative data, and more than 2 years elapsed between 2019 when OFES-CI ratings were provided at the final SCOPE LC and 2021 when qualitative data were coded to obtain gold standard enactment ratings. The present study is also limited by its sample size. A minimum sample size of 30 is generally recommended for calculating ICCs[38] and ICC CIs, unless the ICC is very high, are notably larger with small samples.[39] Although the ICC point estimate (ICC=0.71) suggests that the OFES-CI demonstrates good reliability against our gold standard assessment approach, the true ICC value is somewhere between the reported confidence limits (95% CI=0.46 to 0.86). Additional validation work with larger samples would be valuable and might generate narrower CIs.

The present study focuses on a pragmatic approach to assessing fidelity enactment and it does not require the same rater to provide all assessments (the OFES demonstrated good reliability against a gold standard even though using different raters for some teams generally gives a smaller ICC than using a consistent rater). The present study does not explore the factors that influence fidelity enactment or their mechanism of impact—explorations which could support fidelity enhancement. Fidelity enhancement could be the subject of future research, perhaps by exploring differences between high and low fidelity enactment teams. In-depth study of high and low fidelity teams could also provide insight regarding the OFES-CI's discriminative ability, including its sensitivity and specificity.

Lastly, analogous to Miller's pyramid of competency evaluation where evaluating what someone 'knows' is the lowest level (level 1) and evaluating what someone 'does' is the highest level (level 4),[40] the OFES-CI (as developed in SCOPE) primarily evaluated the extent to which intervention participants 'know how' to enact intervention skills/activities (level 2). OSCEs, where trainees treat a standardised patient, evaluate at level 3 (ie, the trainee 'shows' they have certain skills). Future research could explore ways the OFES-CI approach might build in opportunities for assessing fidelity that sit squarely at level 3. Feasible ways to directly observe fidelity enactment in real-world settings (level 4 of Miller's pyramid) continue to elude researchers.

However, future research should continue to explore creative, feasible ways participants can be observed 'doing' (ie, enacting) intervention skills/activities, perhaps by supplementing a standardised encounter with real-world observation where resources permit.[41]

CONCLUSIONS

The need for robust approaches for assessing implementation/enactment of complex interventions is well documented in the literature. While not a definitive study, our results suggest that the OFES-CI offers a promising, novel and efficient approach for assessing fidelity enactment in QI and other complex interventions. Further use, adaptation and validation of the OFES-CI can enhance understanding of how and why QI and other interventions work, or fail to work, and will contribute knowledge regarding optimal fidelity assessment approaches for complex interventions.

Author affiliations
[1]School of Health Policy and Management, Faculty of Health, York University, Toronto, Ontario, Canada
[2]Faculty of Nursing, University of Alberta, Edmonton, Alberta, Canada
[3]Institute of Health Policy Management and Evaluation, Temerty Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada
[4]Rady Faculty of Health Sciences, University of Manitoba, Winnipeg, Manitoba, Canada
[5]Centre for Care Research, Western Norway University of Applied Sciences, Bergen, Norway
[6]Department of Family Medicine, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada
[7]School of Health and Exercise Science, The University of British Columbia, Kelowna, British Columbia, Canada
[8]School of Kinesiology and Health Science, Faculty of Health, York University, Toronto, Ontario, Canada
[9]Department of Medicine, University of Alberta, Edmonton, Alberta, Canada

Contributors LG led development and testing of the instrument and oversaw data collection and analysis. WB, MD, MH, AW, CR and PGN collected data. MH and WB coded qualitative data. AG facilitated qualitative data analysis. CAE was a lead investigator of the SCOPE study. All reviewed, edited and approved the manuscript. LG is responsible for the overall content as the guarantor.

Funding This study was funded by Canadian Institutes of Health Research (PS 148582).

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval This study involves human participants and was approved by the Research Ethics Boards of the University of Alberta (Pro00000012517), University of British Columbia (H14-03286). Operational approval was obtained from all included facilities as required. SCOPE sponsors and team members were asked for informed consent prior to taking part in the study. Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. De-identified data specific to this study can be requested through the TREC Data Management Committee (joseph.akinlawon@ualberta.ca) on the condition that researchers meet and comply with the TREC and HRDR data confidentiality policies. Data are part of the TREC program of research which has established comprehensive data and intellectual property policies. TREC data are housed in the secure and confidential Health Research Data Repository (HRDR) in the Faculty of Nursing at the University of Alberta (https://www.ualberta.ca/nursing/research/supports-and-services/hrdr), in accordance with the health privacy legislation of participating TREC jurisdictions. The OFES-CI measurement instrument is included in the article.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iDs
Liane Ginsburg http://orcid.org/0000-0002-4436-9198
Matthias Hoben http://orcid.org/0000-0003-3465-315X

REFERENCES

1 Dusenbury L, Brannigan R, Falco M, *et al*. A review of research on fidelity of implementation: implications for drug abuse prevention in school settings. *Health Educ Res* 2003;18:237–56.
2 Marang-van de Mheen PJ, Woodcock T. Grand rounds in methodology: four critical decision points in statistical process control evaluations of quality improvement initiatives. *BMJ Qual Saf* 2023;32:47–54.
3 Moore GF, Audrey S, Barker M, *et al*. Process evaluation of complex interventions: medical research council guidance. *BMJ* 2015;350:h1258.
4 Walton H, Spector A, Tombor I, *et al*. Measures of fidelity of delivery of, and engagement with, complex, face-to-face health behaviour change interventions: a systematic review of measure quality. *Br J Health Psychol* 2017;22:872–903.
5 Skivington K, Matthews L, Simpson SA, *et al*. A new framework for developing and evaluating complex interventions: update of medical research council guidance. *BMJ* 2021;374:n2061.
6 Leis JA, Shojania KG. A primer on PDSA: executing plan–do–study–act cycles in practice, not just in name. *BMJ Qual Saf* 2017;26:572–7.
7 Reed JE, Card AJ. The problem with plan-do-study-act cycles. *BMJ Qual Saf* 2016;25:147–52.
8 Taylor MJ, McNicholas C, Nicolay C, *et al*. Systematic review of the application of the plan-do-study-act method to improve quality in healthcare. *BMJ Qual Saf* 2014;23:290–8.
9 McNicholas C, Lennox L, Woodcock T, *et al*. Evolving quality improvement support strategies to improve plan-do-study-act cycle fidelity: a retrospective mixed-methods study. *BMJ Qual Saf* 2019;28:356–65.
10 Bellg AJ, Borrelli B, Resnick B, *et al*. Enhancing treatment fidelity in health behavior change studies: best practices and recommendations from the NIH behavior change consortium. *Health Psychol* 2004;23:443–51.

11 Lichstein KL, Riedel BW, Grieve R. Fair tests of clinical trials: a treatment implementation model. *Adv Behav Res Therapy* 1994;16:1–29.

12 Walton H, Spector A, Williamson M, *et al*. Developing quality fidelity and engagement measures for complex health interventions. *Br J Health Psychol* 2020;25:39–60.

13 Sprange K, Mountain G, Craig C. Evaluation of intervention fidelity of a complex psychosocial intervention lifestyle matters: a randomised controlled trial. *BMJ Open* 2021;11:e043478.

14 Hasson H. Systematic evaluation of implementation fidelity of complex interventions in health and social care. *Implement Sci* 2010;5:67.

15 Toomey E, Hardeman W, Hankonen N, *et al*. Focusing on Fidelity: narrative review and recommendations for improving intervention fidelity within trials of health behaviour change interventions. *Health Psychol Behav Med* 2020;8:132–51.

16 Proctor E, Silmere H, Raghavan R, *et al*. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. *Adm Policy Ment Health* 2011;38:65–76.

17 Schoenwald SK, Garland AF, Chapman JE, *et al*. Toward the effective and efficient measurement of implementation fidelity. *Adm Policy Ment Health* 2011;38:32–43.

18 Breitenstein SM, Gross D, Garvey CA, *et al*. Implementation fidelity in community-based interventions. *Res Nurs Health* 2010;33:164–73.

19 Toomey E, Matthews J, Guerin S, *et al*. Development of a feasible implementation fidelity protocol within a complex physical therapy-led self-management intervention. *Phys Ther* 2016;96:1287–98.

20 Hoben M, Ginsburg LR, Easterbrook A, *et al*. Comparing effects of two higher intensity feedback interventions with simple feedback on improving staff communication in nursing homes - the INFORM cluster-randomized controlled trial. *Implement Sci* 2020;15:75.

21 Ginsburg LR, Hoben M, Easterbrook A, *et al*. Examining fidelity in the INFORM trial: a complex team-based behavioral intervention. *Implement Sci* 2020;15:78.

22 Ginsburg LR, Tregunno D, Norton PG, *et al*. Development and testing of an objective structured clinical exam (OSCE) to assess socio-cultural dimensions of patient safety competency. *BMJ Qual Saf* 2015;24:188–94.

23 Wagg A, Hoben M, Ginsburg L, *et al*. Safer care for older persons in (residential) environments (SCOPE): a pragmatic controlled trial of a care aide-led quality improvement intervention. *Implement Sci* 2023;18:9.

24 Regehr G, MacRae H, Reznick RK, *et al*. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Acad Med* 1998;73:993–7.

25 van der Vleuten CPM, Schuwirth LWT. Assessing professional competence: from methods to programmes. *Med Educ* 2005;39:309–17.

26 Hodges B, Regehr G, McNaughton N, *et al*. OSCE checklists do not capture increasing levels of expertise. *Academic Medicine* 1999;74:1129–34.

27 Couturier J, Kimber M, Barwick M, *et al*. Assessing fidelity to family-based treatment: an exploratory examination of expert, therapist, parent, and peer ratings. *J Eat Disord* 2021;9:12.

28 Ginsburg LR, Easterbrook A, Massie A, *et al*. Building a program theory of implementation using process evaluation of a complex quality improvement trial in nursing homes. *Gerontologist* 2023:gnad064.

29 Kilo CM. A framework for collaborative improvement: lessons from the institute for healthcare improvement's breakthrough series. *Qual Manag Health Care* 1998;6:1–13.

30 Ginsburg LR, Hoben M, Easterbrook A, *et al*. Fidelity is not easy! challenges and guidelines for assessing fidelity in complex interventions. *Trials* 2021;22:372.

31 Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol Assess* 1994;6:284–90.

32 Krippendorff K. *Content analysis: an introduction to its methodology. 2nd ed*. Thousand Oaks: SAGE Publications, Inc, 2004.

33 Hallgren KA. Computing inter-Rater reliability for observational data: an overview and Tutorial. *Tutor Quant Methods Psychol* 2012;8:23–34.

34 McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychological Methods* 1996;1:30–46.

35 Brown CH, Curran G, Palinkas LA, *et al*. An overview of research and evaluation designs for dissemination and implementation. *Annu Rev Public Health* 2017;38:1–22.

36 McGee D, Lorencatto F, Matvienko-Sikar K, *et al*. Surveying knowledge, practice and attitudes towards intervention fidelity within trials of complex healthcare interventions 11 medical and health sciences 1117 public health and health services. *Trials* 2018;19.

37 Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys Ther* 2005;85:257–68.

38 Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med* 2016;15:155–63.

39 Liljequist D, Elfving B, Skavberg Roaldsen K. Intraclass correlation – A discussion and demonstration of basic features. *PLoS One* 2019;14:e0219854.

40 Miller GE. The assessment of clinical skills/competence/performance. *Acad Med* 1990;65:S63–7.

41 Malau-Aduli BS, Jones K, Saad S, *et al*. Has the OSCE met its final demise? rebalancing clinical assessment approaches in the peri-pandemic world. *Front Med* 2022;9.

42 Kitson A, Harvey G, McCormack B. Enabling the implementation of evidence based practice: a conceptual framework. *Qual Health Care* 1998;7:149–58.

43 Rycroft-Malone J, Harvey G, Kitson A, *et al*. Getting evidence into practice: ingredients for change. *Nurs Stand* 2002;16:38–43.

44 Poss JW, Jutan NM, Hirdes JP, *et al*. A review of evidence on the reliability and validity of minimum data set. *Healthc Manage Forum* 2008;21:33–9.