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

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 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.

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 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), 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, 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. 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. 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. The need for efficient, high-quality, practical approaches to assessment of fidelity enactment has been highlighted by several recent reviews as has the need for studies that outline specific procedures for assessing fidelity. Building on our previous work, 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 (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 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. 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.

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. 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. 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 Model29 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)
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
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 and process evaluation are published elsewhere.

Development of the OFES-CI

The OFES-CI was developed alongside SCOPE following steps outlined by Walton and colleagues for developing high-quality fidelity measures. We also adhered to practices used in our previous work on fidelity assessment and on the use of expert raters. 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

**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

<table>
<thead>
<tr>
<th>Rating Scale: Team’s level of ENACTMENT of SCOPE Activities</th>
<th>NO / VERY LOW ENACTMENT of SCOPE activities appropri. for LC4</th>
<th>BORDERSHINE UNSATISFACTORY ENACTMENT for this stage of SCOPE (LC4)</th>
<th>BORDERSHINE SATISFACTORY ENACTMENT for this stage of SCOPE (LC4)</th>
<th>SATISFACTORY ENACTMENT for this stage of SCOPE (LC4)</th>
<th>VERY HIGH ENACTMENT extensive implementation of SCOPE activities for LC4</th>
<th>UNABLE TO ASSESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 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.

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 intra-class correlation coefficient (ICC) (appropriate when the same pair of raters is not used for all teams) and was found to be good (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 fidelity9 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 data28 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 analysis32 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.

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

<table>
<thead>
<tr>
<th>Data collection approach</th>
<th>Purpose</th>
<th>Data used to arrive at the gold standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>QA Diaries</td>
<td>QAs completed a diary entry after each interaction with a team. Diary entries included both facts (eg, ‘Team tested X change’, ‘% 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.</td>
<td>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.</td>
</tr>
<tr>
<td>Learning Congress Exit Surveys</td>
<td>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 aides led, team accomplishments, support received, and implementation facilitators and barriers. Surveys were completed at each learning congress by care aides, leaders and QAs.</td>
<td>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.</td>
</tr>
<tr>
<td>Observations</td>
<td>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.</td>
<td>Two researchers observed two different activities for each team at LC4: presentation of team storyboards and PDSA progress presentation.</td>
</tr>
</tbody>
</table>

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

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

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) 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 (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).
[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. 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 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-CI rating given. The OFES-CI addresses the practicality problem in fidelity assessment by virtue of its suitable implementation qualities (acceptability, ease of completion, low burden, low training requirements).

Similar to Walton’s findings, 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.

Figure 2 Gold standard fidelity enactment rating and OFES-CI rating difference scores. OFES-CI, Overall Fidelity Enactment Scale for Complex Interventions.
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 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.66 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 (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;37 (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 al3 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.
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. 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—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 and ICC CIs, unless the ICC is very high, are notably larger with small samples. 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-CI 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), 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.

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.
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.44

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
1School of Health Policy and Management, Faculty of Health, York University, Toronto, Ontario, Canada
2Faculty of Nursing, University of Alberta, Edmonton, Alberta, Canada
3Institute of Health Policy Management and Evaluation, Temerty Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada
4Rady Faculty of Health Sciences, University of Manitoba, Winnipeg, Manitoba, Canada
5Centre for Care Research, Western Norway University of Applied Sciences, Bergen, Norway
6Department of Family Medicine, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada
7School of Health and Exercise Science, The University of British Columbia, Kelowna, British Columbia, Canada
8School of Kinesiology and Health Science, Faculty of Health, York University, Toronto, Ontario, Canada
9Department 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 CR and PGN collected data. MH and WB coded qualitative data. LG led development and testing of the instrument and oversaw data collection and 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 148382).

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.akinlayon@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.

Open access
This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.

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

REFERENCES


Online Supplementary Appendix
OFES-CI Rater Training materials and Detailed Instructions to Raters

This appendix includes information on the content, components, and rater training materials and outlines the process for completing the fidelity rating sheets at the final learning congress. There are three sections in this Appendix.

Section 1 – Material for initial rater training that took place prior to Learning Congress 2 (LC2)

Section 2 – Material for additional rater training that took place prior to Learning Congress 4 (LC4)

Section 3 – The final LC4 OFES-CI Package for Raters (including instructions for raters and the actual rating form)
Section 1 - Material for initial rater training that took place prior to Learning Congress 2 (LC2)

LC2 Rater Training Agenda (held via Zoom)

ATTENDEES: All 6 faculty expert raters who will be attending LC2

Meeting AGENDA:
1.1. Review of OSCE-style Fidelity rating form
   • We will review the Rating form, including definitions of fidelity enactment, guidelines to rating and ‘look fors’ and answer any clarification questions

1.2. What constitutes fidelity enactment and what level of is appropriate for LC2?
   • We will discuss the range of implementation / enactment we can expect to see at LC2, recall that teams are now three months into SCOPE and do not have huge amounts of time to devote to the project.

1.3. Review fidelity related concepts of ‘core’ and ‘peripheral’ components and ‘drift’ (in the context of the longstanding fidelity-adaptation debate).
   • How should we take into account ‘drift’ and modification of ‘core’ vs ‘peripheral’ components of SCOPE as we rate fidelity enactment? E.g., the ability to adapt certain, usually more peripheral, aspects of an intervention like SCOPE is critical for successful implementation as it permits some degree of tailoring to ensure suitability to local context. We seek to assess fidelity to the ‘core components’ of SCOPE; discuss fact that pragmatically this may not always be clear.

1.4. Calibration debriefing (see meeting preparation instructions)

PREPARATION for Zoom Training:
PLEASE READ the following CAREFULLY and allow yourself 45 minutes to prepare for the Call by completing the Quantitative Calibration exercise.

You are all familiar with aims of SCOPE. At LC2 all SCOPE teams will do a presentation of their progress so far. You will be asked to rate these LC2 presentations using a quantitative fidelity enactment rating form. We have prepared a calibration exercise for each of us to complete on our own and then discuss during our rater training meeting next week. The calibration exercise uses a video of students who use the model for improvement to improve operational aspects of a heart failure clinic. Their presentation is similar to the presentation SCOPE teams will do at LC2. To complete the calibration exercise please do the following before next week’s training meeting:

   a. Carefully review the GUIDELINES for Rating fidelity ENACTMENT in SCOPE and the rating form available here
   b. Read the 1-page heart failure case backgrounder that goes with the video available here
   c. Watch and score the 8-minute HF improvement project video presentation using the rating form (attachment 1). [Videolink disabled for privacy]
Section 2 – Material for additional rater training that took place prior to Learning Congress 4 (LC4)

LC4 Rater Training (held via Zoom)

ATTENDEES: 4/5 faculty expert raters who will be attending LC4

Meeting AGENDA:
1. Brief review of updated Fidelity rating form
   - Recap changes made to the rating form and procedure following LC3 (Q&A added to presentations, addition of adjustment score and 0.5 rating, role of comment box)
2. What level of fidelity enactment is appropriate for LC4?
   - We will discuss the range of implementation / enactment we can expect to see at the end of SCOPE (LC4), one year after project start; time constraints and other priorities continue to challenge for most SCOPE teams but there is a level of fidelity enactment we are still expecting to see
3. Calibration scenario discussion

PREPARATION for Zoom Training:
Please review the rater instructions and other information in the section 3 (below), including the final LC Fidelity Enactment Rating Form prior to the training meeting allow yourself 20-30 minutes to prepare for the Call.
Section 3 – The final OFES-CI Package for Raters (including instructions for raters and the rating form)

The final LC4 OFES-CI Package for Raters

SCOPE Fidelity Enactment Rater Instructions
1. As an expert regarding the core components of SCOPE, you are being asked to complete the attached sheet by providing a rating of SCOPE fidelity enactment for EACH SCOPE TEAM presentation you are assigned to observe at LC4.
2. Each faculty expert will be assigned 3-4 teams and will do the following three things before completing the fidelity rating sheet:
   2.1. review each of your assigned team’s story boards during the morning activity prior to team presentations;
   2.2. listen to each team deliver their SCOPE progress presentation;
   2.3. listen to Q&A following each presentation and ask a question of the team during the Q&A. There are typically few questions posed to teams so it is helpful all around if you ask a question (i.e. it can furnish additional information that can help you come up with fidelity ratings).

What constitutes Fidelity Enactment in SCOPE?
3. Fidelity enactment is defined in the “Guidelines to Rating” and refers to a team’s actual implementation of the following SCOPE activities:
   3.1. Defining project aims (though this should have been clearly done by previous LCs)
   3.2. Generating change ideas
   3.3. Using PDSA cycles and Measurement to test changes
   3.4. Modifying unsuccessful changes / spreading successful changes to residents/staff across the unit.
   3.5. SCOPE activities are intended to be carried out by teams led by healthcare aides.
   3.6. 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).
4. At LC4 (the end of the SCOPE intervention) raters should be looking for evidence of all of the activities in 3.1-3.5; the following “look fors” are things you would see for teams you are rating as a 3 or 4 on the rating scale:
   4.1. Team generated change ideas relevant to their project aims
   4.2. Team used systematic approach to measurement appropriate to aims, gathered data and plotted measurements over time
   4.3. Team tested change(s), studies the changes, and adjusts or spreads to other residents/staff on the unit.

Calibration
5. Two forms of calibration are important to consider:
   5.1. After you have completed ratings for all of the teams assigned to you, you are encouraged to assess whether you were too lenient or too tough with any of the teams you rated relative to the other teams you rated.
   5.2. When there are multiple raters engaged in the process, calibration helps ensure different raters have a similar perspective on what constitutes low- or high-fidelity enactment. At LC2 disagreements between raters centered around how ‘lenient’ we were in our assessments of team’s measurement and PDSA activities (two raters were slightly hawkish while two were
dovish). Similarly, we noted that in LC3 some raters adhered more rigidly to what constitutes measurement and change in IHI’s Model for Improvement.

6. The following 3 Calibration Scenarios relevant to 5.2 will be discussed during training:

6.1. A team makes a change but it is NOT an appropriate change activity (e.g., it does not relate to the Aim statement or to the clinical area they are working on) - this would get an enactment score of? [Answer not provided prior to discussion: likely a 0]

6.2. A team has done a good job with PDSA but the manager did all the measurement work or all of the change and measurement work. [Answer not provided prior to discussion: this enactment score might be “2-Borderline Satisfactory” because the SCOPE activities were not truly team-based or Healthcare aide-led.]

6.3. A team was engaged in first aspects of the PDSA cycle with healthcare aides pretty fluent defining solid aim to reduce pain by 50% with applicable residents and outlining reasonable change ideas. However, graphing of the measurement data on pain care aids collected was left with the facility leader who didn’t deliver, missing this crucial aspect of feedback in the PDSA cycle. They stated their plan to spread to other units but this failed to materialize in a meaningful way – sharing a few stories of how to track pain may have occurred to other units). [Answer not provided prior to discussion: this would get an enactment rating of around 1 “borderline unsatisfactory enactment for LC4”].

6.4. *Raters are reminded to use the comment box on the rating form if you come across a tricky scenario and feel the need to clarify why you provided a certain rating.

7. Rating Form Completion Reminders

7.1. While the rating scales range from 0-4, it is acceptable to provide a .5 rating (between 2 categories).

7.2. It is important to rate each team you are assigned. However, if you truly ‘cannot assess’ a team, be sure to provide a rationale in the open-ended comments box on the rating form.

7.3. Use bullets 1-4 above and the ‘look fors’ as a guide to what a team in category 3 or 4 on the rating scale would be doing.

7.4. For each rating, indicate either “I would leave this rating as is”; “I might raise it ½ or 1 category”; “I might lower it ½ or 1 cat” (and you can provide notes in the comment box).

7.5. Use the comments box if you wish to explain any of your ratings.
GUIDELINES for Rating scales to assess FIDELITY ENACTMENT in SCOPE
Learning Congress 4

Fidelity Enactment refers to a team’s actual implementation of SCOPE activities* as intended by the intervention (e.g. activities are carried out without so much ‘drift’ that they no longer resemble activities intended by SCOPE).

*SCOPE activities = defining project aims, generating change ideas, using PDSA cycles and Measurement to test changes, spreading successful changes to residents/staff across the unit

The following is intended to provide observers with (1) broad descriptions of enactment rating area and its boundaries; and (2) “Look fors” that, if present, reflect categories 3&4 on the rating scale.

<table>
<thead>
<tr>
<th>GUIDELINE TO FIDELITY ENACTMENT RATING SCALE</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO / VERY LOW ENACTMENT of SCOPE activities appropriate for LC4 / inappropriate activities implement</td>
</tr>
<tr>
<td>0</td>
</tr>
</tbody>
</table>

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

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

– Complete reverse side of this page for EACH TEAM you are assigned –
LC4 Fidelity Rating Form

Observer Initials: _____ Date: _____

[Prepopulate with team name here]

- “Look fors” are intended to prompt you while you assess a team;
- “Look fors” are NOT an exhaustive list of what you should use in arriving at your rating for each scale.
- “Look fors”, if present, would reflect categories 3&4 on the rating scale.

<table>
<thead>
<tr>
<th>ENACTMENT</th>
<th>NO / VERY LOW ENACTMENT of SCOPE activities appropriate for LC4 OR inappropriate activities implemented</th>
<th>BORDERLINE UNSATISFACTORY ENACTMENT for LC4</th>
<th>BORDERLINE SATISFACTORY ENACTMENT for LC4</th>
<th>SATISFACTORY ENACTMENT for LC4</th>
<th>VERY HIGH ENACTMENT extensive implementation of SCOPE activities for LC4</th>
<th>UNABLE TO ASSESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rating Scale: Team’s level of ENACTMENT of SCOPE Activities</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>9</td>
</tr>
</tbody>
</table>

Look fors:
- 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

Please describe any uncertainties regarding your rating of this team: