Development and validation of an A3 problem-solving assessment tool and self-instructional package for teachers of quality improvement in healthcare

Jennifer S Myers ,[1] Jeanne M Kin,[2] John E Billi,[3,4,5] Kathleen G Burke,[6,7] Richard Van Harrison[8]

For numbered affiliations see end of article.

Correspondence to
Dr. Jennifer S Myers, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA;
jennifer.myers2@pennmedicine.upenn.edu

Received 28 July 2020
Revised 20 February 2021
Accepted 10 March 2021
Published Online First 26 March 2021

► http://dx.doi.org/10.1136/bmjqs-2021-013251

To cite: Myers JS, Kin JM, Billi JE, *et al*. *BMJ Qual Saf* 2022;**31**:287–296.

ABSTRACT

Purpose A3 problem solving is part of the Lean management approach to quality improvement (QI). However, few tools are available to assess A3 problem-solving skills. The authors sought to develop an assessment tool for problem-solving A3s with an accompanying self-instruction package and to test agreement in assessments made by individuals who teach A3 problem solving.

Methods After reviewing relevant literature, the authors developed an A3 assessment tool and self-instruction package over five improvement cycles. Lean experts and individuals from two institutions with QI proficiency and experience teaching QI provided iterative feedback on the materials. Tests of inter-rater agreement were conducted in cycles 3, 4 and 5. The final assessment tool was tested in a study involving 12 raters assessing 23 items on six A3s that were modified to enable testing a range of scores.

Results The intraclass correlation coefficient (ICC) for overall assessment of an A3 (rater's mean on 23 items per A3 compared across 12 raters and 6 A3s) was 0.89 (95% CI 0.75 to 0.98), indicating excellent reliability. For the 20 items with appreciable variation in scores across A3s, ICCs ranged from 0.41 to 0.97, indicating fair to excellent reliability. Raters from two institutions scored items similarly (mean ratings of 2.10 and 2.13, p=0.57). Physicians provided marginally higher ratings than QI professionals (mean ratings of 2.17 and 2.00, p=0.003). Raters averaged completing the self-instruction package in 1.5 hours, then rated six A3s in 2.0 hours.

Conclusion This study provides evidence of the reliability of a tool to assess healthcare QI project proposals that use the A3 problem-solving approach. The tool also demonstrated evidence of measurement, content and construct validity. QI educators and practitioners can use the free online materials to assess learners' A3s, provide formative and summative feedback on QI project proposals and enhance their teaching.

BACKGROUND

Improving the quality of healthcare is a universal goal for healthcare practitioners and administrators. A3 problem solving is a structured approach to continuous quality improvement (QI) first employed by Toyota and now widely used by healthcare practitioners and organisations that have adopted the Lean thinking approach to improvement.[1–4] Key elements include understanding the reason for action, defining the current state and performance gap, setting a goal, identifying root causes, choosing countermeasures, formulating action plans and establishing a follow-up plan to measure results. QI efforts are more likely to succeed when these elements are employed.

QI is now a required competency for medical students, residents, practising physicians, nurses, pharmacists and other healthcare professionals worldwide.[5–10] A common approach to developing QI skills involves participation in a QI project (QIP) designed around a gap in local healthcare quality. The use of A3 problem solving as an instructional framework for QI skill development has been described in manufacturing and more recently in healthcare.[11–13] Instruction may occur in formal courses or informally in work settings. While numerous experiential QI curricula have been described, few skills-based assessment tools are available.[14–16] None of the existing QIP assessment tools is specific to the A3 problem-solving approach, nor do they provide an easily replicable method to train educators to assess A3 skills.[17–19]

We combined efforts at our two academic healthcare centres to develop an A3 assessment tool and test its reliability through a series of iterative development

Table 1 Self-instruction package for A3 assessment tool: components and descriptions

Components	Description
1. Instructions for assessing problem-solving A3s (proposal stage)	2-page document that explains the purpose (to improve the development of QI project proposals), introduces the other items in the package, explains how to learn to use the assessment tool and provides some practical tips in performing assessments.
2. A3 template	1-page document that lists the most important content of a proposal A3, illustrates the layout and presentation of information and illustrates some relevant QI tools.
3. A3 content guide	5-page document that includes (1) the purpose and use of A3 problem solving; (2) a description of each A3 section: title, background, current situation, problem statement, goal, analysis, countermeasures, action plan and follow-up plan; (3) a list of resources for A3 problem solving.
4. A3 assessment tool	23-item assessment tool divided into 7 A3s sections. Each item has a 4-point rating scale that includes descriptive anchors. Each section has a space for written feedback. The tool also includes 10 additional items for raters who are familiar with the local context of the QI project being rated.
5. Description of ratings	8-page document that reproduces the A3 assessment tool and for each item includes descriptions of the four levels of rating anchors. (The rating anchors have been incorporated into the assessment tool and appear when a cursor hovers over a rating option.)
6. Learning examples for practice and feedback: ► 3 proposal A3s ► A3 assessment tools to complete ► A3 ratings and their explanations	Individuals learning to assess proposal A3s use these materials to try out performing assessments and receive feedback on their performance. The first A3 is exemplary, with an accompanying set of ratings and explanations of why this A3 content illustrates the highest ratings. The second and third proposal A3s have various deficiencies that result in many items having lower ratings. Learners complete an A3 assessment tool for an A3. Then learners receive immediate feedback by checking their ratings and reasoning with the provided ratings and explanations for various levels of ratings on items.

The A3 template is shown in figure 1. All other materials are included in online supplemental digital content.
QI, quality improvement.

cycles. In order for the A3 assessment tool to be easily learnt and widely used, we wanted to develop and test the assessment tool as the central component of a self-instruction package in learning to assess A3s reliably. Development would necessarily include exploring raters' experiences in using the assessment tool and self-instruction package. Ultimately, the resulting A3 assessment tool and self-instruction package should guide QI educators in assessing learners' A3s, provide consistent formative and summative feedback on QIP proposals and teach A3 problem solving.

METHODS

Development cycles for an A3 assessment tool and self-instruction package

We developed an A3 assessment tool and a self-instruction package to assess proposal A3s as part of their QI teaching or advising and to enhance teaching A3 problem solving (online supplemental digital content). Components of the self-instruction package are described in table 1. The five development cycles for the assessment tool and self-instruction package are summarised in the top of table 2. In each cycle, we sought feedback from our raters. In cycles 3–5, we formally tested inter-rater agreement. We used feedback and reliability performance on items at the end of one cycle to refine concepts, improve language precision and enhance presentation of information during the next cycle. Examples of changes across cycles are presented in the bottom of table 2.

We began the first development cycle in 2017 by working with biomedical and business librarians, who performed a systematic literature search using the keywords "A3 thinking", "A3 problem solving" and "A3 template". They searched eight databases covering health sciences, business and engineering (PubMed, Embase, Cochrane Library, Scopus, Web of Science, Compendex, ABI and Business Sources Complete) and publication types (eg, white papers) produced outside of traditional academic publishing channels. We found only one other example of an A3 assessment tool in the engineering literature,[11] and noted that several types of A3s exist, reflecting the stage of improvement work.[2] We focused on a *problem-solving A3* because our institutions currently teach developing them to analyse a QI problem and propose interventions. A problem-solving A3 includes all the dimensions of problem investigation (background, current state, problem statement, goal, analysis), then proposes recommendations (countermeasures, action plan, follow-up plan) based on the findings. We refer to a problem-solving A3 as simply an 'A3' throughout this paper.

The next step in cycle 1 was to create initial drafts of the A3 template, content guide and assessment tool. We reviewed commonly used A3 templates including ones in use at our institutions.[1–3] We created an A3 template that included key sections of A3s with elements described more clearly and operationally than in existing templates. The content guide provided additional descriptive information and illustrations. The assessment tool addressed each element in the template and characteristics across sections. Each item in the assessment tool has response options that range from 0 to 3. General verbal anchors for the options are 0=not addressed, 1=unclear, 2=general and 3=specific, with phrasing modified to reflect an item's content. We realised that items differed in the information that needed to be assessed. The initial assessment

Table 2 Development of an A3 assessment tool and self-instruction package for QI project proposals: (a) overview of five cycles and (b) examples of adjustments between cycles

(a) Overview of five cycles

Activity	Cycle #1 Summer 2017–Spring 2018	Cycle #2 Spring 2018–Summer 2018	Cycle #3 Summer 2018–Fall 2018	Cycle #4 Fall 2018–Spring 2019	Cycle #5 Spring 2019–Fall 2019
Development and revisions	Literature review Created initial A3 materials ► Template ► Content guide ► Assessment tool shared with A3 teachers for comments	Revised materials Added instructions for use of the self-instruction package and assessment tool	Revised materials	Revised materials Added: ► Description of rating options ► Exemplary and deficient A3 examples with rating explanations	Revised materials Added another deficient A3 example with rating explanations Added automated functions to assessment tool
Checks	Feedback from two raters who assessed one A3	Feedback from two experts who reviewed materials	Test of agreement for 4 raters×4 A3s and rater feedback	Test of agreement for 12 raters×6 A3s and rater feedback	Final test of agreement for 12 raters×6 A3s and rater feedback

(b) Examples of adjustments between cycles

Document	Cycle #1 to cycle #2	Cycle #2 to cycle #3	Cycle #3 to cycle #4	Cycle #4 to cycle #5
A3 template	*Within section*. Removed question: 'What residual issues can be anticipated?'	*Across sections*. Moved analysis section to after goal section to match original order used by Toyota.	*Within section*. Added prompt: 'What is contributing to the problem?'	*Within section*. Added question: 'What will be monitored, by whom, when?'
A3 content guide	(No adjustments)	*Within section*. Added illustration of criteria matrix to countermeasures.	*Within section*. Elaborated: 'process map use' and 'strength of countermeasures'.	*Across sections*. Graphics changed to similar set of colours.
A3 assessment tool	*Across sections*. Better visual distinction between items ratable from A3 only or require context knowledge	*Within section*. Eliminated vague question ('How often is information clearly conveyed in each section of the A3?').	*Within section*. Wording improvement: from 'Are timeframes identified …' to 'Are completing dates identified …'	*Within section*. Two items re-categorised from 'ratable from A3 only' to 'requires contextual knowledge'.

Part (a) of this table provides an overview of each development cycle, including when initial versions of documents were developed and the checks performed at the end of each cycle. We created the A3 template, A3 content guide and A3 assessment tool during the first cycle. Part (b) of this table provides examples of adjustments to these documents that were based on comments and testing at the end of one cycle and incorporated in the next cycle. Documents are listed in hierarchal order, with an adjustment to a document often resulting in parallel adjustments (not shown) to subsequently listed documents. In each cycle, minor wording changes (not shown) were made to the documents to improve clarity of language.

tool had 27 items that could be answered directly from information in an A3 document (eg, How specific is the goal?) and 7 items that required additional knowledge of the local problem context (eg, extent to which important root causes are identified). We decided that individuals unfamiliar with the problem context need only rate items that can be determined from the A3 alone. An experienced QI trainer at each institution reviewed and used the materials, then provided feedback.

Cycle 2 incorporated feedback from cycle 1. Then two external Lean experts reviewed the materials with two of the authors (JEB, JMK). In cycle 3, suggestions from the experts were incorporated and formal tests of agreement began. Each test included raters from our two academic healthcare centres. Four individuals (two physicians with QI teaching experience and two non-physician QI professionals) rated four A3s. Their feedback and performance indicated that agreement in assessments would be enhanced through more detailed definitions and guided experience in applying them. In cycle 4, we added a 'description of ratings' document that elaborated operational definitions of individual rating options. We also added examples of exemplary and deficient A3s with rating explanations and the opportunity to assess an A3 and compare ratings against a standard for immediate feedback on performance. The test of agreement expanded the number of raters from 4 to 12 and the number of A3s from 4 to 6. In cycle 5, we added another deficient A3 with rating explanations to compare against a standard. Automated functions were added to the assessment tool to facilitate referencing definitions and totaling scores.

In cycles 3 through 5, we developed exemplary and deficient A3 training examples and A3s used to test inter-rater agreement. First, the authors (JSM, JMK) reviewed examples of A3s submitted by learners in QI methods courses for healthcare professionals (eg,

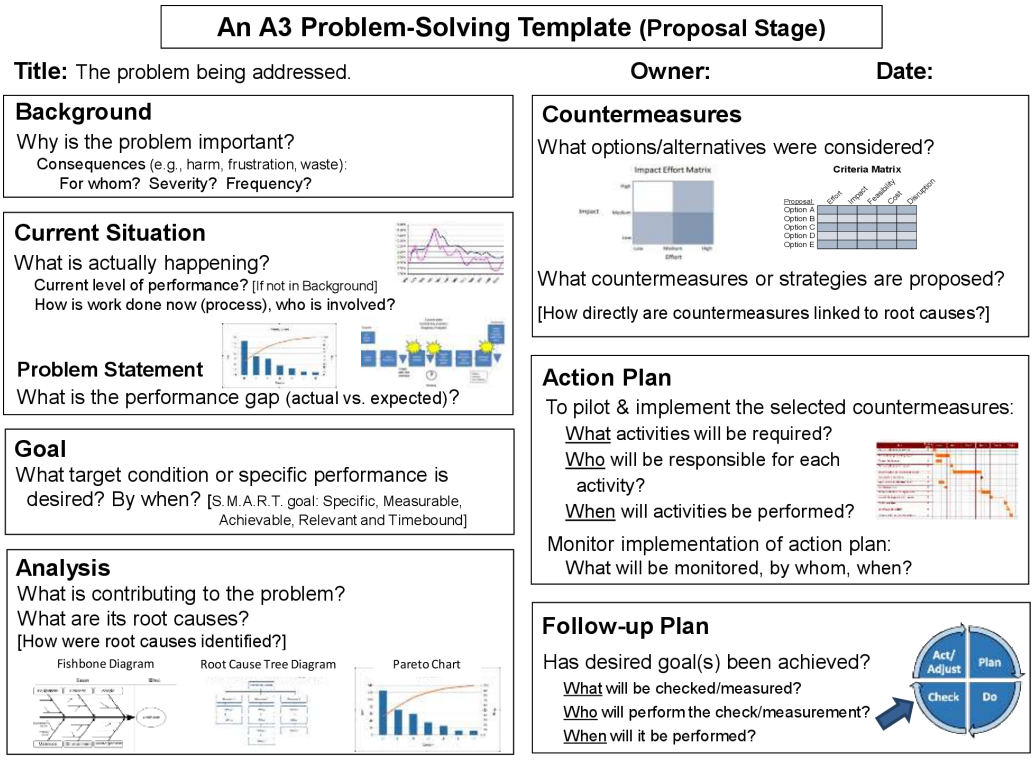

Figure 1 An A3 problem-solving template (proposal stage). Toyota developed the A3 document. This A3 proposal template was modified from previously published versions[1–3] and variations used by Lean educators at our institutions.

physicians, nurses, other healthcare team members) in training (eg, medical students, residents, graduate nursing students) at our institutions. We used course evaluations of A3s to identify examples of excellent, good and poor A3s. Then, we modified most of the A3s by improving some elements (eg, adding completion dates for action plan items) and making other elements worse (eg, adding a countermeasure that did not correspond to a listed root cause) to provide a range on items across the A3s. The three training A3s addressed evidence-based treatment for epilepsy, patient congestion in a clinic and improving the accessibility of cardiac catheterisation films. The six A3s assessed in cycle 5 addressed patient throughput in a psychiatric emergency room (ER), time to decision-making for chest pain patients in the ER, access to care for patients with diabetes after renal transplant, unnecessary phlebotomy in the hospital and equipment waste in the operating room.

Check on cycle 5 of the assessment tool and self-instruction package

Cycle 5 was the culmination of our work. Its check had two objectives: (1) assess inter-rater agreement among raters using the assessment tool and self-instruction package and (2) learn about the raters' experiences and views in using the self-instruction package and performing assessments.

The final A3 template is presented in figure 1. The final A3 assessment tool (online supplemental digital content) has 23 items that can be assessed from the A3 document itself and an additional 10 items that require knowledge of the local context.

Our sample size to test inter-rater agreement was based on practical feasibility for the number of raters and the number of A3s assessed.[20] We felt that 4 hours was the maximum time commitment that we could reasonably request of volunteer raters. Cycle 4 demonstrated that raters could go through the self-instruction package and rate six A3s in approximately 4 hours. We recruited 12 raters for cycle 5 knowing that the increased number of raters would increase precision in estimating inter-rater agreement. The design of 12 raters rating 23 items on 6 A3s produced 72 ratings per item and 1656 ratings overall.

We identified 12 individuals from our two academic healthcare centres (6 from each) and invited them by email to participate as raters. All raters were at least proficient in QI. We selected raters with some, but varying QI teaching experience to reflect the types of individuals most commonly involved in teaching QI in healthcare. Four raters were non-physician QI professionals who routinely led QI initiatives and taught QI as part of their work. The other eight raters were physicians with experience teaching and/or advising students, residents and fellows in QIP work. Four of the eight had been teaching QI for >2 years while the other four had been teaching QI for <2 years.

One of the authors (JSM, JMK, RVH) had a 10 min phone conversation with each rater, orienting the

individual to the study and confirming access to the online self-instruction materials. Raters had 1 month to complete the self-instruction package, rate the six A3s, and submit their ratings.

We created a structured feedback form and distributed it to raters at the time of the orientation phone call (see online supplemental digital content, last section). The form had 19 open-ended items addressing: study orientation, the self-instruction package, the A3 assessment tool and their overall experience with the tool and self-instruction package. Raters provided written feedback when they submitted their A3 ratings and participated in a short debriefing phone call led by one of the investigators. During the call raters could clarify and elaborate upon their comments.

Analysis

We used intraclass correlation coefficients (ICCs) as the primary method to quantify inter-rater agreement. The three variables are rater, A3 and item rating. Values range from 0 to 1. The value is 1 if raters give similar ratings (low variation) to an item within an A3, but ratings differ (high variation) between A3s. The value is 0 if ratings vary within an A3 item as much as they vary between A3s. While guidelines for interpreting ICCs vary, a frequently quoted interpretation is: <0.40 is poor, 0.40–0.59 is fair, 0.60–0.74 is good and 0.75–1.0 is excellent.[21] Lower ICCs reflect greater variation in ratings for an A3 item, so as ICC values decrease the width of an ICC's CIs increases. For our design of 12 raters and 6 A3s, examples of the decreasing precision (95% CI) with which an ICC is measured for an item are: 0.90 (0.77–0.98, within 'excellent'), 0.75 (0.44–0.95, 'fair' to 'excellent') and 0.50 (0.23–0.87, 'poor' to 'excellent').

We calculated ICCs for each of the 23 rating items. To reflect a rater's overall assessment of an individual A3, for each A3 we calculated each rater's mean assessment on the 23 items. A rater's mean rating for an A3 was treated as an additional item for which the ICC was calculated. The 95% CIs for ICCs were also calculated. The ICCs and CIs were calculated using 'R' software for statistical computing based on a single rater, absolute agreement, two-way random effects model.[22]

The ICC is less appropriate as a measure of inter-rater agreement when ratings are similar across A3s. Little variation in ratings within an A3 is similar to the little variation between A3s, resulting in an artificially low ICC, even though raters actually agree and provide similar rating values for an item on all of the A3s. To check that a limited range of scores on an item across A3s might methodologically lower an ICC, we first calculated within each of the six A3s an item's mean score over the 12 raters. Then, we used the means for an item across the six A3s to calculate across the six A3s the overall item mean and the SD of item means. A low SD for an item mean across the six A3s indicates a limited range (little variation) in scores between A3s. For these items, we reviewed the actual scores across A3s to confirm that raters agreed in providing similar rating values across A3s.

In addition to analysing the raters' assessments of items on A3s, we collated qualitative information from raters' feedback forms and debriefing calls and reviewed responses for illustrative themes.

RESULTS

The ICCs and 95% CIs for agreement over a range of scores for the 12 raters across the six A3s are shown in table 3 for the overall A3 rating and the ratings for each of the 23 individual items.

For overall A3 assessment (mean of ratings on an A3's 23 items), the ICC is 0.89 (95% CI 0.75 to 0.98), indicating excellent reliability across raters over a range of scores. For individual items, the ICCs for 17 items ranged from 0.57 to 0.97, indicating fair to excellent reliability; the ICCs for three items (#2, #16, #17) ranged from 0.41 to 0.46, indicating marginally fair reliability.

For the remaining three items (#1, #11, #14), the ICCs range from 0.10 to 0.39, suggesting poor reliability across a range of scores. However, these items did not have a wide range of scores. As shown in table 3, these three items have the lowest SDs (0.28 to 0.55) of the 23 items. For these items, raters generally agreed on the items' scores, but the scores were similar across the six A3s. For example, for item #11 with an ICC of 0.10, with possible ratings ranging from 0 to 3, the means of the 12 rating scores on each of six A3s were 2.9, 2.9, 2.8, 2.7, 2.6 and 2.2. While the raters highly agreed in rating this item between A3s, the variability of scores across A3s was insufficient to demonstrate agreement across a range of scores using an ICC. For items #1, #11 and #14, the lack of variation across A3s methodologically lowered ICCs, limiting our ability to confirm agreement across a range of scores. However, the low SD for these items demonstrate substantial agreement on the score among raters on the items across the six A3s.

For the 20 items with more variation across A3s, the items with higher ICCs tend to have simpler content that focuses on only one element of the A3. For example, the item with the highest ICC is #20. 'Are estimated completion dates identified for each action item (ie, 'when')?' (ICC=0.97). In contrast, items with ICCs in the 'fair' inter-rater agreement range (ICCs 0.40–0.59) require raters to relate multiple elements of information simultaneously, for example, item #17. 'How many of the proposed countermeasures are linked to identified root causes?' (ICC=0.46).

The six raters from each of the two institutions used the rating scales similarly (mean ratings of 2.10 and 2.13, p=0.57). Across institutions, the eight physicians provided slightly higher ratings than the four QI professionals (mean ratings of 2.17 and 2.00,

Table 3 Inter-rater agreement (intraclass correlation coefficients) on overall mean score and individual item scores

Item	Intraclass correlation		For the mean score of an item on an A3 (mean of 12 raters)	
	Coefficient	95% CI	Mean across 6 A3s*	SD across 6 A3s
Overall assessment of A3s (mean of 23 item scores†)	0.89	0.75 to 0.98	2.1	0.51
Individual items				
Background *Why is the problem important?*				
1. Negative consequences (eg, harm, frustration, waste): how specific is the clearest statement of a negative consequence of the problem?	0.32	0.11 to 0.77	2.7	0.37
2. Individuals/Groups impacted by the negative consequences (eg, harm, frustration, waste): how specific is the clearest statement identifying an impacted individual, group/unit or organisation?	0.44	0.19 to 0.84	2.5	0.61
3. Severity of the negative consequences (eg, harm, frustration, waste): how specific is the clearest statement of the severity (eg, extent/amount) of at least one negative consequence?	0.71	0.45 to 0.94	2.3	0.82
4. Frequency of the negative consequences (eg, harm, frustration, waste): how specific is the clearest statement of the frequency (# events/unit of time) of at least one negative consequence?	0.68	0.41 to 0.93	1.8	1.01
Current situation *What is actually happening?*				
5. Current level of performance	0.71	0.46 to 0.94	1.8	0.90
6. How is work done (process/workflow)?	0.72	0.47 to 0.94	1.8	1.07
7. Clear identification of who is involved in performing the work?	0.71	0.45 to 0.94	1.5	1.01
8. Performance problem/gap?	0.58	0.31 to 0.90	1.8	0.90
Goal *What target condition or specific performance is desired? By when?*				
9. How specific is the goal?	0.79	0.57 to 0.96	2.0	0.83
10. Is the goal measurable?	0.60	0.33 to 0.91	2.3	0.68
11. How relevant is the goal to addressing the problem?	0.10	0.0 to 0.52	2.7	0.28
12. How time-bound (clear timeframe for accomplishment) is the goal?	0.96	0.90 to 0.99	1.9	1.49
Analysis *What is contributing to the problem? What are its root causes?*				
13. Is the display of method(s) for analysing root causes easy to understand? (eg, fishbone diagram, '5-whys'/root cause tree diagram, Pareto chart)	0.65	0.38 to 0.92	2.1	0.91
14. How clear are the identified root causes?	0.39	0.15 to 0.81	2.3	0.55
Countermeasures *What options/alternatives were considered? What countermeasures/strategies are proposed?*				
15. How many options for countermeasures were considered?	0.78	0.55 to 0.96	2.7	0.60
16. Identify the strongest countermeasure considered. How strong is it?	0.41	0.17 to 0.82	2.1	0.55
17. How many of the proposed countermeasures are linked to identified root causes?	0.46	0.21 to 0.85	2.0	0.85
Action plan *To pilot and implement the selected countermeasures: what, who, when?*				
18. For the action plan on the A3, how clearly are activities described (ie, 'what' is to be done)?	0.60	0.33 to 0.91	2.3	0.68
19. Are individuals identified to be responsible for each action item to be carried out (ie, 'who')?	0.90	0.77 to 0.98	2.4	1.14
20. Are estimated completion dates identified for each action item (ie, 'when')?	0.97	0.93–1.0	2.5	1.18
21. Is monitoring planned for the implementation of actions (what will be monitored, by whom, when)?	0.57	0.30 to 0.89	1.3	1.06
Follow-up plans *Checking whether desired goal(s) was achieved?*				
22. Is follow-up planned to measure achievement of the desired goal(s) (what will be measured, by whom, when)?	0.83	0.63 to 0.97	1.7	1.00
Across A3 sections				

Continued

Table 3 Continued

Item	Intraclass correlation		For the mean score of an item on an A3 (mean of 12 raters)	
	Coefficient	95% CI	Mean across 6 A3s*	SD across 6 A3s
23. How clearly does the title identify the problem to be addressed?	0.56	0.29 to 0.89	2.3	0.60

Each item has response options that range from 0 to 3 on a 4-point scale. Each response option has verbal anchors appropriate for the item, for example, 0=not addressed, 1=vague, 2=somewhat specific and 3=very specific. The response anchors for each item and their illustrative descriptions and comparisons are presented in the 'Description of Ratings' in the online supplemental digital content.
For each of 6 problem-solving A3s, 12 raters assessed each of 23 items. This produced a total of 1656 ratings, including 12 ratings for each item on each A3, 72 ratings per item across the 6 A3s and 276 ratings per A3 across items.
*The six A3s used to assess inter-rater agreement were modified to increase the range of scores across A3s on several items. The mean scores along with their SD help indicate the extent of variation across A3s for the item. The mean scores do not necessarily reflect a representative sample of student's scores.
†The overall assessment of an A3 is the mean of the 12 raters' assessments for each of the 23 items on an A3 (276 ratings).

p=0.003), but the small difference is not practically meaningful.

On the feedback forms, raters reported that the work took an average of 3.5 hours: the self-instruction package took 1.5 hours (range 1.0–3.0 hours) and rating the six A3s took 2.0 hours (range 1.0–3.5 hours). Illustrative comments about their learning and rating experience are presented in table 4. Overall, raters reported that the self-instruction package and assessment tool were easy to learn and worthwhile to use. For example, "I thought it was easy. I think this tool is going to be a great way to set expectations and give feedback about student A3s". One rater noted "but [I] had to make sure I wasn't inferring information and only evaluated what was on the A3".

DISCUSSION

This study developed and demonstrated the reliability of a tool to assess the quality of learners' investigations and recommendations for QI problems in healthcare using the A3 approach. The assessment tool was developed as part of a self-instruction package to assist a broad range of educators in efficiently learning how to reliably assess and provide feedback on learners' A3 documents. We found that 12 raters using the assessment tool and self-instruction package could reliably rate items across six A3s, with excellent agreement across raters over a range of scores on the overall rating of an A3 and with fair to excellent agreement on 20 items. For the remaining three items, raters agreed in item scoring, but the limited range of scores across A3s precluded confirming agreement across a range of scores. Ratings were similar for raters from different institutions and functionally similar for physician and QI professional raters. The self-instruction package allowed raters to learn to use the assessment tool in about 1.5 hours. Raters found the package and tool easy to learn and worthwhile to use.

Table 4 Illustrative feedback from raters on the A3 self-instruction package and assessment tool

Topic	Responses
A3 template	'The one-page template was really, really well-done in terms of having all the information there especially for people who are learning it for the first time'.
Practice assessing A3s	"Extremely helpful. I appreciated the explanations for why different scores were selected". "I found [the practice] incredibly helpful in providing a systematic and comprehensive way to review the A3s. We all have our focuses and particular areas of expertise/interest, and the standard ratings helped mitigate my personal biases about which aspects to provide feedback on". '….it is a lot of reading. May consider other types of learners and how that information could be packaged for audio/visual learners'.
Applying the assessment tool	'It is a brilliant and pragmatic tool. It was also enjoyable (fun) to use'. "I thought it was easy. I think this tool is going to be a great way to set expectations and give feedback about student A3s". "It was easy in that it confirmed, standardized, and systematized many of the best practices I've learned in my experience doing/teaching process improvement. Everything struck me as an accurate representation of the fundamental concepts". "Yes [I found the assessment tool easy to use], but had to make sure I wasn't inferring information and only evaluated what was on the A3".
Prepare you to better evaluate an A3	'Yes, sharpened understanding and ability to evaluate topics where don't know clinical content as well'. 'Yes. The most helpful components of the package were the description of assessment options, the 'good' A3 example, and the A3 template'.
Will use the package and assessment tool	"I want it right now to use in teaching residents". 'It will be useful to have a consistent tool that's in use across the organization'.

Three other studies reported developing assessment tools for QIP. Leenstra *et al* developed the Quality Improvement Project Assessment Tool (QIPAT-7) in 2007, Rosenbluth *et al* developed the Multi-Domain Assessment of Quality Improvement Projects (MAQIP) in 2017 and Steele *et al* developed the Quality Improvement Project Evaluation Report (QIPER) in 2019.[17–19] Our study adds to this body of literature. Rather than develop a new conceptual framework, we built on the widely recognised Lean A3 problem-solving approach to QI, which an increasing number of healthcare organisations have adopted. For these institutions, our materials facilitate integration of QI operations and QI education for healthcare professionals, educators and learners at all levels. This integration supports high-quality patient care and is now an expectation for healthcare systems that sponsor graduate medical education programmes in the USA.[23] Building on the established A3 framework, we identified specific aspects of A3s to assess and provide educators with a visual template that embeds common QI tools, a companion content guide for the template, examples, practice with feedback and links to resources. Our package of materials is the first to provide training examples of assessments of completed proposals, providing external benchmarks for teachers (and learners). We have gone beyond previous work by demonstrating consistency across raters who are at different institutions, are physicians and QI professionals and are not members of the research team. While we tested the materials on individuals with some experience performing and teaching QI, we anticipate that the self-instruction materials will assist novice QI educators. The assessment tool and instructional package are available online at no cost and require only 2 hours to learn, facilitating their broad use.[24]

The process of developing and testing the reliability of the assessment tool also demonstrated several aspects of its measurement validity—the extent to which it measures what it claims to measure. The first step in establishing content validity was to review the literature on A3 content and templates, assemble and refine the model A3 template and have experts and teachers of A3 problem solving agree that this was the appropriate content to measure. Experts and teachers also agreed that the rating tool represents the content of the A3 template and the logic underlying it. As a component of content validity, 'face' validity is evident in most statements in the template being quoted in items to be rated. Construct validity is demonstrated through items performing in conceptually expected ways, such as items asking about the presence or absence of one element of information being rated more reliably than items involving simultaneous consideration of multiple elements.

Our sequence of development cycles and refinements identified insights that are useful for the QI education and assessment efforts of others. One insight is to distinguish between assessments based on the A3 document alone and assessments based on additional knowledge of the local problem context. Assessments based on the A3 document alone should be consistent among raters. Assessments based on knowledge of the local problem will vary with the assessor's knowledge. Another insight is to help learners differentiate between the QI problem ('what is the specific performance gap') and consequences of the problem ('why the problem is important'). Both learners and raters may use previous knowledge to assume that a problem is important with no explicit statement of why it is important. More precise wording and examples help both learners and raters realise that consequences of a problem are separate from the problem being addressed. Another insight from examining previously developed A3s is that having a plan for monitoring whether the proposed actions are actually implemented ('intervention fidelity') is frequently overlooked.[25] Including this concept in the A3 template and assessment tool helps ensure that this important step is addressed.

Our study has several limitations. The assessment tool does not address actual outcomes of QIPs that have been completed. We focused on the proposal stage because development of well-researched, well-analysed and well-considered proposals for interventions is the foundation for carrying out successful QI efforts. Some healthcare settings may not use the A3 framework on which our materials are based. However, use of the framework is sufficiently widespread that teachers and learners should be aware of this approach to developing QIPs. Including only 6 A3s and 12 raters limited the ranges sampled and ICC precision but reasonable evidence of inter-rater agreement was demonstrated. The generalisability of the results to other settings and professional roles is uncertain. Our raters were from one country and two academic centres, which possibly provided some common contexts regarding views of QI and the QI training available. The tool would likely not perform as well with individuals inexperienced in QI or with no experience teaching QI. However, within groups likely to be responsible for teaching and assessing A3s, the results potentially apply to a range of settings, personnel and training levels because our study included raters from different professions (physicians, QI professionals) with experience ranging from some to extensive proficiency in performing QI and teaching QI, and because the A3s that were the basis for testing agreement were authored by different professional student groups (eg, physicians, nurses, pharmacists). Finally, the raters typically knew one of the authors personally, potentially biasing feedback towards being more favourable. However, in our preliminary cycles, similarly chosen raters provided critical feedback that prompted changes. Since previous feedback included negative comments that were addressed, the

favourable feedback in the final cycle appears to reflect reasonably unbiased views.

The A3 assessment tool and self-instruction package can be used for future research. The effect of being better trained to assess A3s has yet to be explored for subsequent outcomes such as providing better feedback or teaching effectiveness. Also to be explored is the impact of the assessment tool and self-instruction package on the quality of learners' A3s and actual QIP outcomes. Assessments and feedback could be provided prospectively to learners to determine the impact of longitudinal formative feedback on A3s. The materials could also be provided to learners to determine the extent to which learners on their own can improve their A3s and those of peers. Future research could also expand studies of reliability of agreement among raters across institutional settings and individuals with different levels of QI knowledge and skills. Finally, supplementing the documents in the current self-instruction package with materials in video format may enhance learning efficiency and effectiveness.

In summary, this study provides evidence of the reliability and validity of a tool to assess the quality of A3 project proposals in healthcare. The assessment tool was developed as the focus of a self-instruction package to assist a broad range of QI educators and practitioners to assess learners' A3s, to provide consistent formative and summative feedback on QIP proposals and to enhance their teaching of A3 problem solving. We demonstrated that after using the self-instruction package, raters from different institutions and professional backgrounds who are proficient in QI and have some experience teaching QI can reliably assess A3s. Raters performed ratings in about 1.5 hours and found the package and tool to be easy to learn and worthwhile to use. The materials are available on our institutional website at no charge.[24] The minimal investment required to use the materials facilitates their widespread use by individuals teaching QI to healthcare professionals and by individuals performing QI in healthcare.

Author affiliations
[1]Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, USA
[2]Quality, Michigan Medicine, University of Michigan, Ann Arbor, Michigan, USA
[3]Medicine and Learning Health Sciences, Michigan School of Medicine, University of Michigan, Ann Arbor, Michigan, USA
[4]Health Management and Policy, School of Public Health, University of Michigan, Ann Arbor, Michigan, USA
[5]Integrative Systems and Design, College of Engineering, University of Michigan, Ann Arbor, Michigan, USA
[6]Biobehavioral Health, University of Pennsylvania School of Nursing, Philadelphia, Pennsylvania, USA
[7]Leonard Davis Institute of Health Economics, University of Pennsylvania, Philadelphia, Pennsylvania, USA
[8]Learning Health Sciences, University of Michigan Health System, Ann Arbor, Michigan, USA

Acknowledgements The authors would like to thank the following individuals who participated as raters in this study: Amber-Nicole Bird, Ryan Buckley, Debbie Paliani Burke, Caitlin Clancy, Kevin DeHority, Tammy Ellies, Sara Figueroa, Laurel Glaser, Kevin Gregg, Katie Grzyb, Katy Harmes, Jessica Hart, Michael Heung, Elena Huang, Chloe Hill, Christopher Klock, Jamie Lindsay, Erin Lightheart, Rosalyn Maben-Feaster, Patricia Macolino, Neha Patel, Anita Shelgikar, Elizabeth Valentine, Kimberly Volpe, Jason Wagner, Sarah Yentz. The authors would also like to thank Eric Ethington and John Shook, well known Lean thought leaders, who reviewed and provided feedback on an early version of the materials; the librarians Maylene Kefeng Qiu, Mia Wells, Melanie Cedrone and Sherry Morgan who assisted with the literature review and Larry Gruppen, who provided comments on the manuscript draft.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data relevant to the study are included in the article or uploaded as supplementary information.

ORCID iD
Jennifer S Myers http://orcid.org/0000-0002-3450-9572

REFERENCES

1 Shook J. *Managing to learn: using the A3 management process to solve problems, gain agreement, mentor, and lead*. Cambridge, MA: Lean Enterprise Institute, Inc;, 2008.
2 Sobek D, Smalley A. *Understanding A3 thinking: a critical component of Toyota's PDCA management system*. New York, NY: Productivity Press, Taylor & Francis Group, 2008.
3 Jimmerson C. *A3 problem solving for healthcare: a practical method for eliminating waste*. New York, NY: Healthcare Performance Press, 2007.
4 Jimmerson C, Weber D, Sobek DK. Reducing waste and errors: piloting lean principles at intermountain healthcare. *Jt Comm J Qual Patient Saf* 2005;31:249–57.
5 Association for American Medical Colleges. Quality improvement and patient safety competencies across the learning continuum. Available: https://www.aamc.org/data-reports/report/qipscompetencies [Accessed March 1, 2020].
6 Accreditation Council of graduate medical education common program requirements. Available: http://www.acgme.org/What-We-Do/Accreditation/Common-Program [Accessed March 1, 2020].
7 Brown DR, Warren JB, Hyderi A. AAMC core entrustable professional activities for entering residency Entrustment concept group. finding a path to entrustment in undergraduate medical education: a progress report from the AAMC core entrustable professional activities for entering residency Entrustment concept group. *Acad Med* 2017;92:774–9.

8 World Federation for Medical Education. Basic medical education WFME global standards for quality improvement. accessed at. Available: https://wfme.org/download/wfme-global-standards-for-quality-improvement-bme/ [Accessed on July 25, 2020].
9 Quality & Safety Education for Nurses Competencies. Accessed at. Available: https://qsen.org/competencies/pre-licensure-ksas/ [Accessed on July 25, 2020].
10 American College of clinical pharmacists: ACCP clinical pharmacist competencies. accessed at. Available: https://www.accp.com/docs/positions/guidelines/Saseen_et_al-2017-Pharmacotherapy__FINAL [Accessed on July 25, 2020].
11 Amos A, Taylor K, Johnson K. Assessing the quality of the A3 thinking tool for problem solving. In: Ahram TZ, Karwowski W, eds. *Advances in the service side of human engineering. advances in intelligent systems and computing*. Switzerland: Springer International Publishing, 2017: 49–61.
12 Waits SA, Reames BN, Krell RW, *et al*. Development of team action projects in surgery (TAPS): a multilevel team-based approach to teaching quality improvement. *J Surg Educ* 2014;71:166–8.
13 Kim CS, Lukela MP, Parekh VI, *et al*. Teaching internal medicine residents quality improvement and patient safety: a lean thinking approach. *Am J Med Qual* 2010;25:211–7.
14 Wong BM, Etchells EE, Kuper A, *et al*. Teaching quality improvement and patient safety to trainees: a systematic review. *Acad Med* 2010;85:1425–39.
15 Boonyasai RT, Windish DM, Chakraborti C, *et al*. Effectiveness of teaching quality improvement to clinicians: a systematic review. *JAMA* 2007;298:1023–37.
16 Peiris-John R, Selak V, Robb G, *et al*. The state of quality improvement teaching in medical schools: a systematic review. *J Surg Educ* 2020;77:889–904.
17 Leenstra JL, Beckman TJ, Reed DA, *et al*. Validation of a method for assessing resident physicians' quality improvement proposals. *J Gen Intern Med* 2007;22:1330–4.
18 Rosenbluth G, Burman NJ, Ranji SR, *et al*. Development of a multi-domain assessment tool for quality improvement projects. *J Grad Med Educ* 2017;9:473–8.
19 Steele EM, Butcher R, Carluzzo KL, *et al*. Development of a tool to assess trainees' ability to design and conduct quality improvement projects. *Am J Med Qual* 2020;35:125–32.
20 Ionan AC, Polley M-YC, McShane LM, *et al*. Comparison of confidence interval methods for an intra-class correlation coefficient (ICC). *BMC Med Res Methodol* 2014;14:121.
21 Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol Assess* 1994;6:284–90.
22 R Core Team. R: a language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria. URL, 2013. Available: http://www.R-project.org/
23 Co JPT, Weiss KB, CLER Evaluation Committee. CLER pathways to excellence, version 2.0: Executive summary. *J Grad Med Educ* 2019;11:739–41.
24 A3 Problem-Solving Resources – Center for Healthcare Improvement & Patient Safety | University of Pennsylvania Perelman School of Medicine (upenn.edu) (accessed on Dec 24, 2020).
25 Etchells E, Woodcock T. Value of small sample sizes in rapid-cycle quality improvement projects 2: assessing fidelity of implementation for improvement interventions. *BMJ Qual Saf* 2018;27:61–5.