Contents lists available at ScienceDirect

Surgery

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

How physicians change: Multisource feedback driven intervention improves physician leadership and teamwork

Jinwei Hu, MD[a,b,*], Robert Lee, MD[a], Sarah Mullin, MS[a], Steven Schwaitzberg, MD, MS[a,b], Larry Harmon, PhD[c,d], Paul Gregory, PhD[d], Peter L. Elkin, MD[a,e]

[a] Department of Biomedical Informatics, Jacob's School of Medicine and Biomedical Sciences, University at Buffalo, NY
[b] Department of Surgery, Jacobs School of Medicine and Biomedical Sciences, University at Buffalo, NY
[c] Department of Psychiatry & Behavioral Sciences, University of Miami Miller School of Medicine, FL
[d] PULSE 360 Program/Physicians Development Program, Inc, Miami, FL
[e] Buffalo VA Medical Center, VA Western New York Healthcare System, NY

ARTICLE INFO

Article history:
Accepted 2 June 2020
Available online 6 August 2020

ABSTRACT

Background: Multisource feedback provides a method of quantitatively assessing and improving physician professionalism, interpersonal communication, teamwork, and leadership behaviors. We sought to determine whether tiered educational interventions can improve measurements of multisource feedback for physicians across specialties, and whether multisource feedback baseline measurements and improvements after intervention vary by specialty designation.

Methods: Multisource feedback assessments were performed on physicians from academic (34%) and community hospitals (66%) in the United States and Canada. PULSE 360 Survey data was obtained on 1,190 physicians from primary care (25%), surgical (46%), and other (29%) specialties. Physician respondents were 75% male and 24% female. Raters included administrators, colleagues, staff, and self-ratings with an average of 35.7 ratings per physician.

A leadership teamwork index was measured before and after delivery of educational intervention. Three tiers of intervention were used depending on baseline leadership teamwork index score: (1) report only, (2) debriefing only, and (3) debriefing and development.

Results: Surgeons had a significantly lower baseline leadership teamwork index at 59.9, whereas primary care and specialists started with an leadership teamwork index of 67.1 and 65.9, respectively. Those who participated in a tier 3 intervention had the greatest change from an average baseline leadership teamwork index of 36.6 to 56.3 on follow-up. Surgeons experienced the largest mean increase of 9.1 leadership teamwork index points after intervention, whereas medicine specialists had a mean increase of 6.7 leadership teamwork index points.

Conclusion: Baseline multisource feedback scores vary by specialty and improve based on feedback, goal-setting, coaching, and education. In particular, physicians who start with low scores have the greatest potential for leadership teamwork index improvement.

Introduction

At its core, medicine is a profession founded on science, art, ethics, communications, and relationships—whether between a physician and patient, a physician and staff, or a physician and professional colleagues.[1,2] The consequences of failures in communication and interprofessional teamwork are evident across all specialties.[3,4] In the medical intensive care unit, an increased rate of interprofessional collaboration has been shown to lead to improved patient outcomes.[5] At Veterans Affairs Medical Centers, surgical services utilizing a high degree of feedback and care coordination significantly lowered morbidity related to surgical procedures.[6] Furthermore, a recent article by Cooper et al demonstrated that surgeons with higher reports of unprofessional behavior appeared to have higher risks of postoperative complications.[7] Evidence suggests that 52% to 70% of adverse events are caused in part owing to human factors such as teamwork and communication breakdown.[8,9] As a result, physician competency as defined by the Accreditation Council for Graduate Medical

[*] Reprint requests: Jinwei Hu, MD, Department of Surgery, Jacobs School of Medicine and Biomedical Sciences, University at Buffalo, 100 High St, Buffalo, NY 14203.
E-mail address: jinweihu@buffalo.edu (J. Hu).

https://doi.org/10.1016/j.surg.2020.06.008
0039-6060/© 2020 Elsevier Inc. All rights reserved.

Education now includes professionalism and interpersonal communication skills.[10] The modern physician must at the same time be an assertive manager, a compassionate caretaker, and, perhaps above all, an ethical leader.

Initially adopted by the business and management industries, multisource feedback (MSF), or 360-degree feedback review, has been gaining traction as a powerful tool to assess physician competency in the workplace.[11] Anonymous feedback is solicited from multiple sources within the work environment, including colleagues, superiors, subordinates, and patients, thus providing an unbiased view of performance.[12,13] MSF allows for service-line leaders to commend those who score well and target the delivery of education and coaching to those who would benefit the most. Furthermore, MSF provides a mechanism by which to assess and reinforce behavioral change over time.[14]

The results of MSF have been validated across a variety of medical specialties.[15] In pediatric residents, it has been shown to improve physician behavior more than traditional, single-evaluator feedback techniques.[16] It has been correlated with patient satisfaction scores, and those who have scored poorly are at higher risk of malpractice claims.[17,18] In addition, it has been increasingly adopted as part of the recredentialing process in several institutions. The Council of Academic Hospitals in Ontario, Canada has developed guidelines for MSF evaluations and has implemented these across the entire province's healthcare system.[19]

Like any other evaluation tool, the success of MSF depends on program implementation, the nature of the feedback that the participants receive, and how the feedback is translated into action.[20,21] Lockyer identified 4 tenets governing the success of an MSF program—organizational support, steering committee work, monitoring, and psychometric design and testing.[22] Apart from the initiation of an MSF program, the form of intervention and further follow-up is also essential to behavior change.[20,23,24] Physician performance and subsequent improvement is reliant on overcoming barriers such as misunderstanding of roles and responsibilities, perception of areas of expertise, hospital culture, and power dynamics.[25] The complexities of physician feedback and behavioral improvement thus demand a standardized approach for reliability, yet must employ individualized implementation science in order to be effective and transformative.

In this study, we examined the implementation of a multifaceted, proprietary MSF program, using the PULSE 360 Survey, across several large hospital systems encompassing all medical specialties. To our knowledge, this is the largest standardized trial of a MSF program. Physicians were provided with PULSE 360-degree review reports and then stratified into several tiers of normalized leadership-teamwork index (LTI) scores. Based on these tiers, standardized behavioral interventions were performed including reporting only, reporting with 1 individualized telephone debriefing, and both of the above in addition to ongoing individualized telecoaching and online education. Physicians were then retested using the PULSE 360 Survey and provided with their evaluation reports. We report whether the composite LTI score obtained by a physician, which assesses leadership and teamwork effectiveness, improves based on intervention, and whether there are specialties that differ in their LTI scores.

Materials and methods

Setting

From 2002 to 2017, PULSE 360 Survey programs were implemented in over 850 medical facilities including community hospitals, academic medical centers, clinics, and practice groups. The current sample of data includes a subset of physicians who participated in repeat assessments during approximately a 3- to 24-month period. Physicians participated in the PULSE 360 program as part of a professional development activity for a group (eg, surgery department) or as individual participants needing or seeking to garner feedback from colleagues and staff to improve their leadership, teamwork, and communication-related skills and behaviors. All physicians received feedback on a version of the PULSE 360 Survey, which is a proprietary web-based survey system (PULSE 360 Program, Miami, FL) that consists of 20 to 52 questions depending on the version of the survey. The survey consists of Likert-type scaled questions asking raters to assign a score based on the extent to which they have observed a particular behavior from the feedback recipient as well as to type-in their opinions on open-ended comment questions asking about what they would like the physician to start, stop, and keep doing. Raters are identified through a 2-step process. First, the physician selects from an online directory of all potential raters, the colleagues and staff with whom they interact or the leaders who oversee those with whom they interact, to invite them to provide anonymous feedback (generally at least 10 of each type). For the second step, the list of selected raters is automatically emailed to the physician's validator (often the chief, chair, or medical director who oversees the provider), who reviews the list and may confidentially add any missing raters to help ensure that a fair, representative, and comprehensive sample of raters are invited. Surveys typically run for 21 days with automated reminders to encourage a higher rate of responses. The average response rate for PULSE 360 surveys is about 75%. The average number of raters per physician at baseline was 35.7 with a standard deviation (SD) of 31.1; at follow-up, the average number of raters was 26.5 with a SD of 23.7. Once the feedback reports are prepared, the validator typically approves the results and determines the debriefing and development plan for the feedback recipient.

360-degree review data

All the 1,190 participants (405 from academic hospitals and 785 from community hospitals) participated in a 360-degree review process, which included an initial and a follow-up assessment. One participant did not participate in follow-up assessment. Physician specialties contained 3 broad categories of providers. The physician category of primary care was defined to include family medicine, hospitalists, internal medicine, and pediatrics. Specialists included anesthesiology, dermatology, emergency medicine, internal medicine subspecialties, neurology, psychiatry, pathology, physical medicine and rehabilitation, radiology, and subspecialties. Surgeons included cardiothoracic surgery, general surgery and subspecialties, obstetrics and gynecology and subspecialties, neurosurgery, ophthalmology, orthopedic surgery, otolaryngology, and urology. Participation was voluntary and all 360-degree review data were anonymized by removing distinguishing characteristics to help prevent identification of reviewers (Fig 1).

360-degree intervention

After the debriefing and development planning call with facility leadership, physicians were assigned to 1 of 3 levels of intervention depending on the decisions made by their organization: The physician (1) received his or her PULSE 360 feedback report and no other intervention; (2) received a telephone debriefing in which the PULSE-trained coach reviewed the feedback and helped set "excellence goals" using a proprietary goal-setting interface (PULSE 360 Program), or (3) received the PULSE-coach telephone debriefing, set goals using the standardized activity, participated in ongoing individualized coaching, watched educational videos assigned by the coach based on feedback-identified developmental

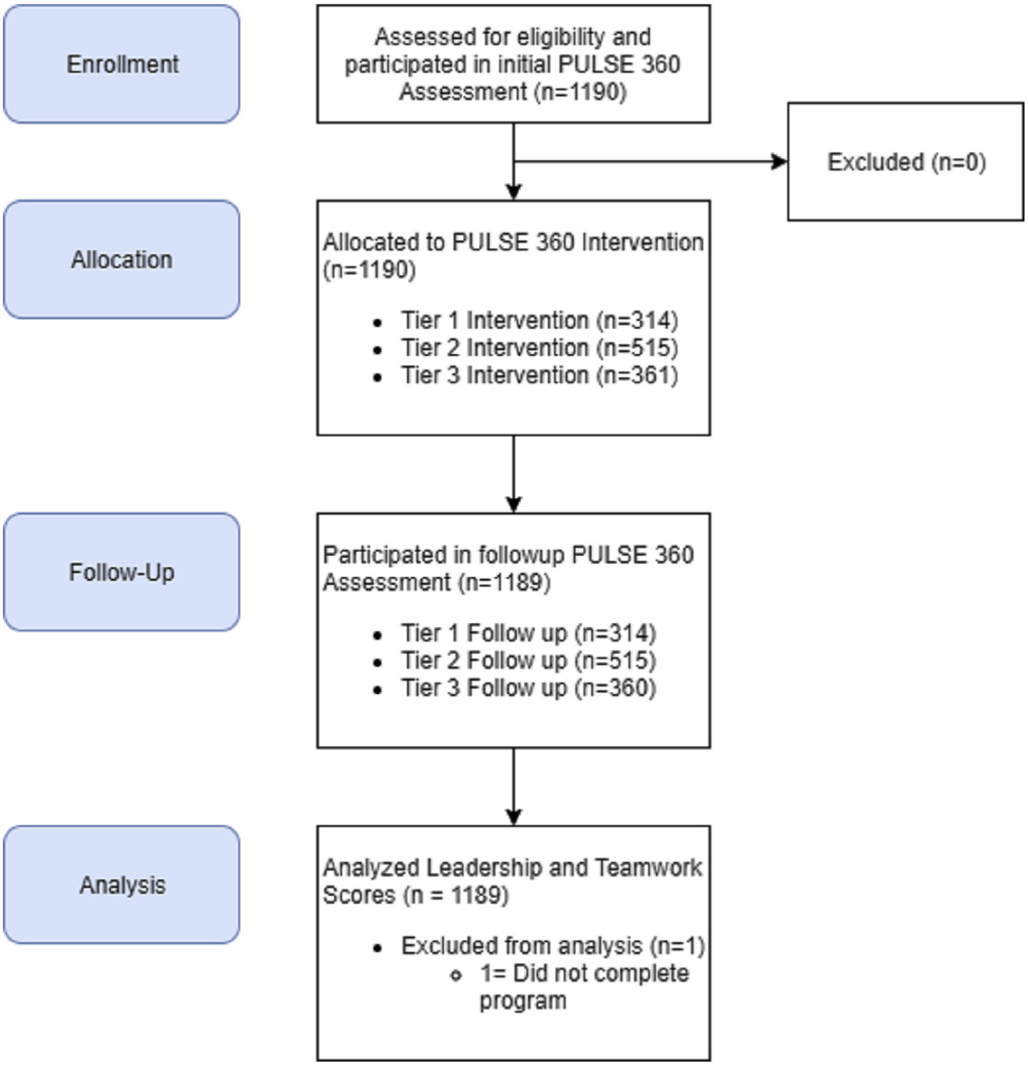
Enrollment
Assessed for eligibility and participated in initial PULSE 360 Assessment (n=1190)
Excluded (n=0)
Allocation
Allocated to PULSE 360 Intervention (n=1190)
Tier 1 Intervention (n=314)
Tier 2 Intervention (n=515)
Tier 3 Intervention (n=361)
Follow-Up
Participated in followup PULSE 360 Assessment (n=1189)
Tier 1 Follow up (n=314)
Tier 2 Follow up (n=515)
Tier 3 Follow up (n=360)
Analysis
Analyzed Leadership and Teamwork Scores (n = 1189)
Excluded from analysis (n=1)
1= Did not complete program

Fig 1. Pulse 360 study. Physicians were selected for participation in the PULSE 360 program by their institutions. One thousand one hundred and ninety physicians participated in the PULSE 360 program, and 1 participant was lost to follow-up and excluded from the study.

needs and completed the video-related learning activities, and discussed how the tools and techniques presented in the educational videos could be applied to the provider's specific work environment and challenges.

The coaches had master's or doctoral degrees in psychology or education, had received structured training on how to interpret 360-degree survey feedback, had overcome physician defensiveness in response to receiving negative feedback, had participated in codebriefing activities similar to shadowing with more experienced coaches, and were rated by physicians on the effectiveness of the coach's debriefing skills.

About 3 to 24 months after the assigned intervention, the physician participated in a follow-up PULSE 360-degree survey, and these results were returned to the participant, though the time interval for this follow-up and returning the feedback may have been longer depending on the healthcare organization's scheduling requests (Fig 2).

LTI

The PULSE 360 LTI score is based on a formula that uses a composite of positive and negative behavioral attributes of the PULSE 360 survey to create a single score. Equal weight is given to both positive and negative attributes, so if the average of all negative attribute scores exceeds that of the positive attribute scores, it is possible for the feedback recipient to have a negative LTI score. The range of scores is from −100 to 100 with the national baseline mean score for physicians equal to 68.9 for example. Most feedback recipients score between 0 and 100, and it is extremely rare for someone to receive a negative score.

Positive attributes, such as "treats other respectfully," and negative attributes, such as "talks down to others," are rated on a 5-point Likert scale from (1) not at all to (5) to a great extent. The number of behavioral items included in the calculation of the LTI varied somewhat depending on the survey questions chosen by the facility. However, internal consistency reliability analyses place the average Cronbach's alpha for the scale at .89 and average inter-rater agreement (intraclass correlation coefficient-1) at .85. LTI scores were averaged across rater groups, including self-assessments, providers, staff, supervisors, and administrative staff, to create a mean LTI score for an initial assessment and a mean LTI score for a follow-up assessment. All other variables in the analysis remained constant across time.

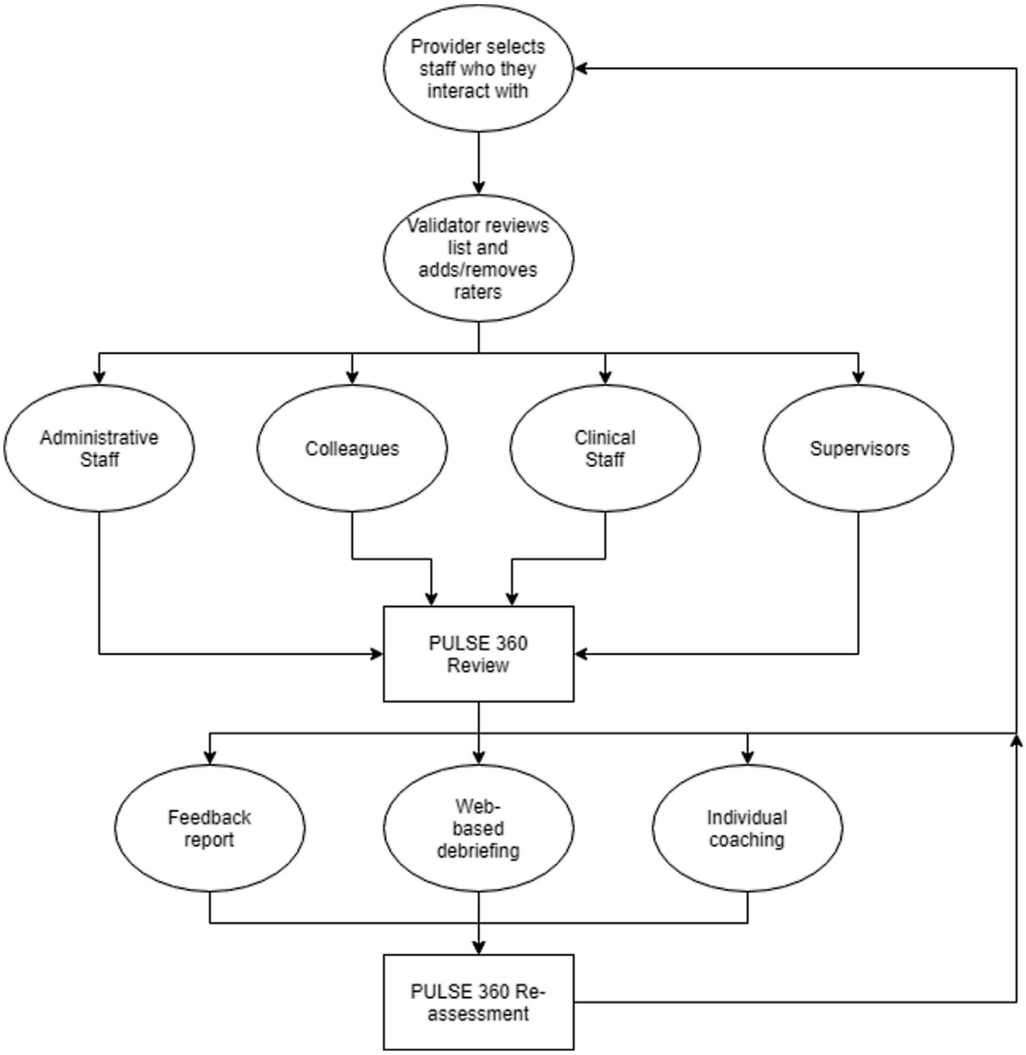

Fig 2. Pulse 360 process. Each physician selected staff they interacted with, a validator added or removed raters, and the physician underwent MSF review from these raters. The physician was then assigned on tiers 1, 2, or 3 intervention based on their ratings. They underwent reassessment and re-evaluation with the PULSE 360 program.

Statistical analysis

Statistical analysis was performed using R 3.5.1 (R Foundation for Statistical Computing, Vienna, Austria) and Minitab 19 (Minitab, LLC, State College, PA). The package geepack (R Foundation for Statistical Computing) was used to estimate quasilikelihood generalized estimating equation (GEE) models to account for the correlation between subjects across evaluations. The GEE model assesses the global LTI score for the subjects at their initial and follow-up time points with fixed effects for the tiered intervention with tier 1 as the reference category, specialty categories with primary care as the reference category, region with Northeast as the reference category, and sex. For reference, GEE estimates may be interpreted similarly to ordinary least square regression coefficients. Subject was used as a random variable to account for correlated outcomes from different time intervals. Since the follow-up survey took place across varying time differences ranging from 3 to 24 months, to account for differing lengths of time between assessments, delay was calculated in days and incorporated in the correlation structure. We investigated the main effects and 2-way and 3-way interactions between our predictors intervention, sex, region, specialty, and time to select a final model. In addition, a second model was run only on individuals with a baseline LTI score less than 52 (25th percentile for the baseline LTI, $n = 380$) to mitigate the natural ceiling effect present for scores at the top of the scale.[26]

In this model, we have assumed exchangeable correlation structure for within subject variance, although an autoregressive order 1 and independence within-subject covariance structures were tested. GEE is robust to mis-specified correlation structures because the Huber-White sandwich variance estimator was used.[27] In addition, coefficient estimates were similar, and the exchangeable correlation had the lowest quasilikelihood under the independence model criterion. We have also assumed that there is a linear relationship between the covariates and the response where the identity link and the Gaussian distribution were assumed. LTI is skewed left (mean = 67.3, median = 74.7, skew = −0.958) and, because GEE does not depend on distributional assumptions, LTI was not transformed for ease of interpretation. To assess pairwise comparisons, Tukey multiple comparison tests were used.

Results

Survey data were obtained on 1,190 physician participants with no missing data; 785 (66%) came from a community or community-associated hospital, and 405 (34%) came from a primarily academic-

Table I
Physician participant and study characteristics

	n (%)
Sex	
Male	904 (0.76)
Female	286 (.24)
Facility	
Academic	405 (0.34)
Community	785 (0.66)
Region	
Northeast	285 (0.24)
Midwest	186 (0.16)
Southeast	147 (0.16)
Southwest	30 (0.025)
West	492 (0.41)
Canada	50 (0.41)
Rater group	
Administrative staff	487 (0.075)
Provider	2,140(0.33)
Self	747 (0.12)
Staff	2,109 (0.33)
Supervisor	871 (0.14)
Uncategorized	113 (.018)

associated hospital. Overall, 904 (76%) were male, while 286 (24%) were female (Table I). The majority of the physicians were from the West and Northeast regions. Out of the 1,190 reviewed, 295 (24.8%) were from a primary care specialty, 560 (46.1%) were from a surgical specialty or subspecialty, and 347 (29.1%) were from a medicine subspecialty or other specialty.

There were on average 35.7 (SD = 31.1) ratings for each participant at the initial assessment and 26.5 ratings (SD = 23.7) ratings at the subsequent assessment. The overall person-averaged LTI scores ranged from −26 to 100 for the initial assessment and −31.5 to 100 for the follow-up assessment with an overall average of 63.4 (SD = 27.7, median = 70.7) for the initial assessment and 71.2 (SD = 22.1, median = 76.7) for the follow-up assessment. LTI scores for surgeons, on average, were lower at baseline and had the greatest on average change compared with primary care and specialists (Fig 3, Table II). The baseline scores were only for physicians who had follow-up surveys and do not represent the average scores for all providers.

There were 314 (27%) participants who had a tier 1 intervention, 515 (44%) participants who underwent a tier 2 intervention, and 361 (31%) who participated in a tier 3 intervention. The average LTI for tier 3 intervention was low at baseline and had the greatest change in LTI at follow-up. For all physicians, those who participated in a tier 3 intervention had the greatest change from an average LTI of 36.62 to 56.28, whereas the participants who underwent a tier 1 or 2 intervention began at similar LTI scores with an insignificant difference in postintervention LTI (Table II, Fig 3).

For the GEE analysis, the highest-level significant interactions were 2-way interactions between specialty and time, intervention and time, and specialty and intervention. The 3-way interaction was removed from the model (Wald statistic = 14.7, $P = .099$). The main effects for the predictors region and sex were also included in the model. The model with the exchangeable correlation structure had a QIC = 14,289 compared with QIC = 14,298 for the autoregressive order 1 correlation structure. With respect to LTI physician scores, specialty, intervention, time, and region, the model was statistically significant compared with the null model (Wald statistic = 1,179, $P < .001$) (Table III).

Post hoc Tukey multiple pairwise comparison tests for intervention by time, averaged over specialty category and region, shows tier 3 is statistically different than tiers 1 and 2 interventions at both baseline and follow-up (Table IV). Tier 2 shows an average increase in 2.4 LTI points from the baseline assessment at follow-up (Z = −4.18, $P < .001$). Tier 3 shows an average increase in 19.4 LTI points from the baseline assessment to follow-up (Z = −20.42, $P < .001$). Tier 1 change from baseline to follow-up was not statistically significant. LTI points from baseline assessment to follow-up for primary care shows an average increase in 6.56 (Z = −8.12, $P < .001$). For specialists, the average increase in LTI points was 7.84 (Z = −9.76, $P < .001$). Finally, the average increase in LTI points for surgeons was 9.39 (Z = −13.78, $P < .001$). For the interaction of specialty category and intervention, other specialists with tier 1 had a mean of 12.8 (Z = 4.72, $P < .001$) more LTI points compared with surgeons with the same tiered intervention. For tier 3, primary care specialists had a mean of 9.7 (Z = 3.84, $P = .004$) more LTI points compared with surgeons. Multiple comparison tests for region, averaged over intervention, specialty category, and time, shows that the West region is statistically different from the Northeast and Southeast regions (Z = 5.97, $P < .001$, Z = 5.37, $P < .001$, respectively) with the West region having a higher LTI. The Midwest had a higher LTI than the Northeast and Southeast (Z = 3.03, $P = .03$, 3.69, $P = .003$, respectively).

For the second model incorporating only participants with low baseline LTI, the most parsimonious model contained only the 2-way interaction of intervention and time (Wald statistic = 10.4, $P = .73$, compared with the model presented in Table III). For this model, average baseline LTI was similar (tier 1: mean = 33.6, SD = 18.2, tier 2: mean = 33.7, SD = 13.9, tier 3: mean = 26.5, SD = 16.6), allowing better comparison and an opportunity to have similar increases without a ceiling effect present. This model remained consistent with the overall model, with a statistically significant intercept effect of −7.06 (standard error = 2.74, $P < .001$) for tier 3 compared with tier 1. The slope effect for tier 3 shows a mean increase of 13.44 LTI points compared with tier 1 (standard error = 3.53, $P < .001$).

The specialties in this sample with the bottom baseline quartile of LTI scores were neurosurgery, cardiology, orthopedic surgery, general surgery, and obstetrics and gynecology, respectively 51.20, 55.19, 56.24, 56.74, and 58.42. The specialties with the top baseline quartile of LTI scores were emergency medicine, anesthesiology, ophthalmology, otolaryngology, and radiation oncology, respectively 72.94, 69.5, 69.11, 68.80, and 68.65 (Table V).

Out of the 54 behavioral characteristics polled, for all physicians, the lowest rated 5 teamwork behavioral characteristics, from lowest to highest, were related to “complains about mistakes respectfully,” “motivates hard work,” “makes requests respectfully,” “focused under stress,” and “approachable when stressed.” For all physicians, the highest rated negative teamwork behavioral characteristics, from highest to lowest, were “reduced others’ job satisfaction,” “criticizes indirectly,” “snaps at others,” “uses offensive gestures,” and “intimidates others” (Table VI).

Discussion

To our knowledge, this is the first large study demonstrating the efficacy of multisource, feedback-guided interventions. Behavioral change is difficult to effect. We have shown that using a tiered system of web-based reports, debriefing, goal-setting, and individualized coaching can improve physician leadership, teamwork, and communication skills.

Across all specialties, physicians with a low baseline LTI score benefited more significantly from tier 3 intervention (debriefing, goal-setting, individualized coaching, and feedback-linked educational videos) and tier 2 intervention (debriefing alone) compared with tier 1 intervention (no interventions other than receiving a feedback report). Region differences showed higher LTI’s for the West and Midwest when compared with the Northeast and

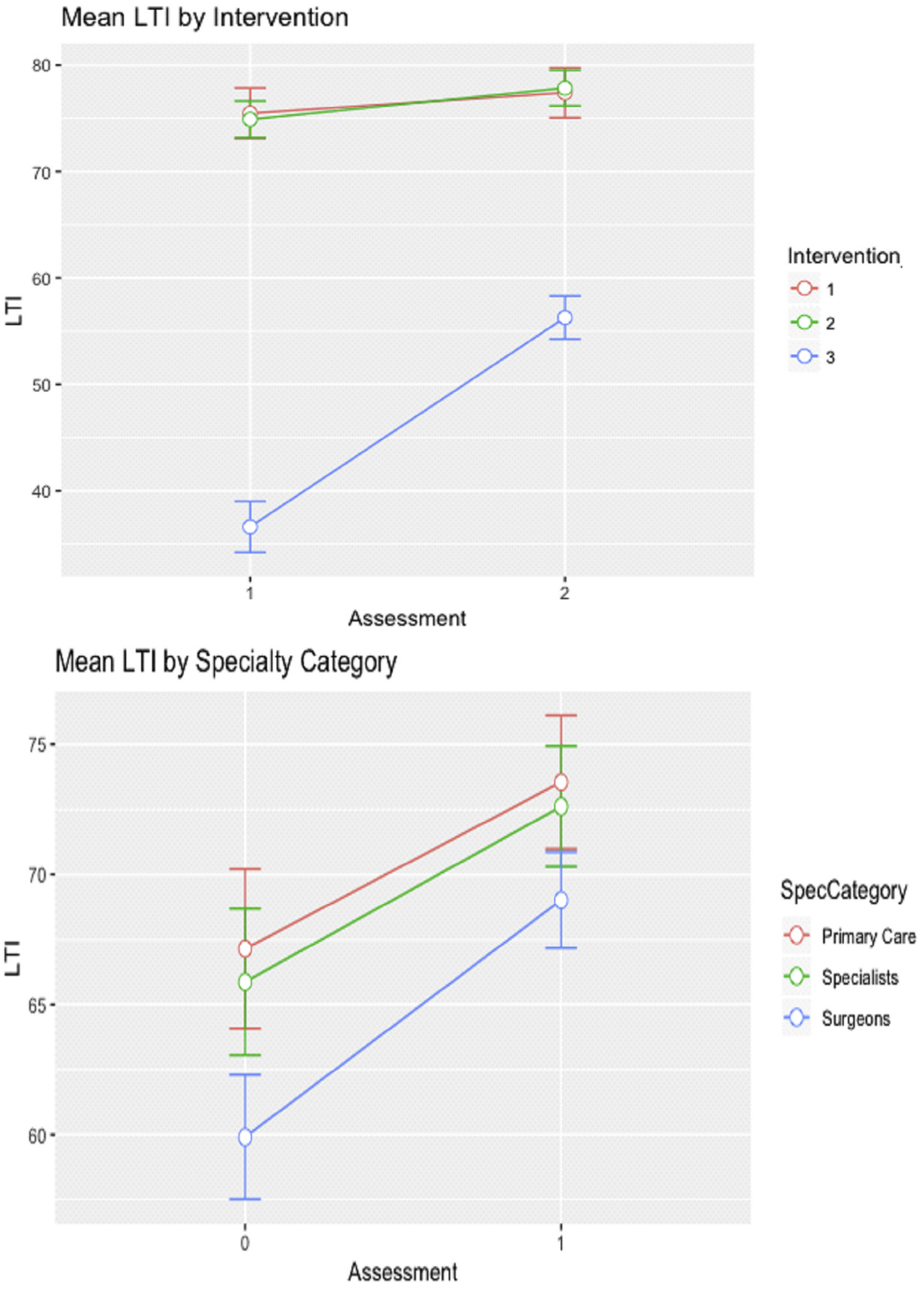

Fig 3. Mean LTI by intervention and specialty. Mean LTI before and after intervention were compared with respect to intervention and specialty. Those who underwent tier 3 intervention improved the most. Surgical specialists had the lowest baseline LTI scores with the most improvement.

Southeast. There were no sex associations with the change in LTI scores. The program was successful in improving the teamwork, leadership, and interpersonal capabilities of the participants regardless of sex. Therefore, this study suggests that MSF reviews demonstrate a positive effect in physicians with low scores.

In the business and management professions, coaching has been demonstrated to be an effective strategy in transforming both leaders and cultures.[23,27,28] In previous studies, coaching has also been effective in developing physician leaders in orthopedic surgery.[14] In our study, those who underwent tier 1 intervention (feedback report) improved their LTI scores from 75.5 to 77.4, a difference of 1.9, and those who participated in tier 2 intervention (debriefing alone) improved their LTI scores from 74.9 to 77.8, a difference of 2.9. However, the most impressive improvements were seen with those who participated in tier 3 intervention (individualized coaching and feedback-linked, physician-related

Table II
Leadership-teamwork index by intervention and specialty

Intervention	Time	*n* (%)	LTI (mean)	SD	CI
1	Initial	314 (26.4)	75.5	21.4	2.38
1	Follow-up	314 (26.4)	77.4	21.1	2.34
2	Initial	515 (43.3)	74.9	19.9	1.73
2	Follow-up	515 (43.3)	77.8	19.3	1.67
3	Initial	361 (30.3)	36.6	23.2	2.41
3	Follow-up	360 (30.3)	56.3	19.6	2.03
Specialty category					
Primary care	Initial	295 (24.8)	67.1	26.7	3.06
Primary care	Follow-up	295 (24.8)	73.6	22.4	2.57
Specialists	Initial	347 (29.1)	65.9	26.6	2.82
Specialists	Follow-up	347 (29.2)	72.6	21.9	2.34
Surgeons	Initial	548 (46.1)	59.9	28.6	2.37
Surgeons	Follow-up	547 (46.0)	69.0	22	1.83

CI, confidence interval.

Table III
GEE model coefficients for predicting LTI change

	Estimate (β)	Sandwich SE	Wald test statistic	*P* value
Intercept	70.028	2.913	577.98	<.001
Intervention tier 2	2.929	2.715	1.16	.2806
Intervention tier 3	−28.881	3.085	87.66	<.001
Time	0.639	0.963	0.44	.5069
Specialists	3.761	2.549	2.18	.14
Surgeons	−9.848	3.046	10.45	.0012
Midwest	6.773	2.236	9.17	.0025
Southeast	−1.219	2.403	0.26	.6118
Southwest	1.257	3.46	0.13	.7164
West	9.35	1.88	24.74	<.001
Canada	2.614	3.722	0.49	.4826
Male	−1.418	1.306	1.18	.2774
Intervention (2) × time	0.383	0.994	0.15	.7
Intervention (3) × time	17.384	1.24	196.55	<.001
Intervention (2) × specialists	−8.506	3.507	5.88	0.0153
Intervention (3) × specialists	-9.559	3.724	6.59	0.0103
Intervention (2) × surgeons	10.579	3.916	7.3	0.0069
Intervention (3) × surgeons	−1.296	3.918	0.11	.7408
Specialists x time	1.281	1.13	1.29	.2569
Surgeons x time	2.832	1.067	7.04	.008

SE, standard error.

Table IV
Multiple pair-wise comparisons using Tukey for intervention by time

Contrast	Estimated LTI	SE	Z	*P* value
1, baseline - 2, baseline	−3.6	1.603	−2.26	.211
1, baseline - 3, baseline	32.5	1.879	17.29	<.0001
1, baseline - 1, follow-up	−2	0.779	−2.58	.102
1, baseline - 2, follow-up	−4.4	1.606	−3.74	.003
1, baseline - 3, follow-up	13.1	1.781	7.36	<.0001
2, baseline - 3, baseline	36.1	1.72	21	<.0001
2, baseline - 1, follow-up	1.6	1.579	1.02	.911
2, baseline - 2, follow-up	−2.4	0.572	−4.18	<.0001
2, baseline - 3, follow-up	16.7	1.61	10.39	<.0001
3, baseline - 1, follow-up	−34.5	1.861	−18.54	<.0001
3, baseline - 2, follow-up	−38.5	1.714	−22.47	<.0001
3, baseline - 3, follow-up	−19.4	0.95	−20.42	<.0001
1, follow-up - 2, follow-up	−4	1.599	−2.5	.123
1, follow-up - 3, follow-up	15.1	1.77	8.54	<.0001
2, follow-up - 3, follow-up	19.1	1.588	12.04	<.0001

Results are averaged over the levels of specialty category, region, and sex, Tukey method *P* value adjustment for a family of 6 estimates. Contrasts are defined as (intervention, time) where (1, baseline - 2, baseline) is interpreted as intervention tier 1 compared with intervention tier 2 at baseline.
SE, standard error.

educational videos), with an improvement of 36.6 to 56.3, a difference of 19.7—a significantly positive result compared with the less interactive tiers 1 and 2 interventions. This further supports that ongoing systematic, directed coaching and education may be necessary to effect greater behavioral change.

We found that there were specialty-specific differences in starting LTI scores and how each specialty improved after intervention. The GEE model demonstrated that, while all specialties increase their LTI with intervention, surgeons increased their LTI the most. Furthermore, those surgical specialists who underwent tier 1 intervention had a significantly lower LTI than other specialists who underwent tier 1 intervention. Surgical specialists were found to have lower starting LTI scores at 59.9, that with directed intervention, were raised to 69.0—which is near the baseline starting score for primary care physicians and medical specialists (respectively, 67.1 and 66.0). Furthermore, primary care physicians and medical specialists did not have as large an improvement in LTI scores compared with surgeons.

Surgical specialties have a unique subset of care they must provide that often involves acute, critical clinical scenarios where directive rather than empowering behaviors are encouraged. It is conceivable that surgical specialists lower LTI scores are owing to differences in their training paradigms, level of burnout, and relationships. Indeed, there is a body of literature describing the unique relationships that must be recognized among the surgical specialties, such as the surgeon-anesthesiologist relationship and dynamics in the operating room in order to provide safe and effective care.[29,30] It is clear that among the cohort of surgical

Table V
Baseline LTI scores for all subspecialties

Specialty	Baseline LTI mean	SD	Number of physicians rated
Neurosurgery	51.20	38.05	22
Cardiology	55.19	27.29	41
Orthopedic surgery	56.24	30.30	96
General surgery	56.74	28.75	137
Obstetrics & gynecology	58.42	29.03	93
Pediatrics	61.45	27.86	38
Cardiothoracic surgery	61.76	19.52	37
Not boarded (general practice)	62.84	36.19	20
Neurology	63.32	25.14	42
Plastic surgery	64.91	31.28	22
Urology	65.19	33.70	32
Radiology	67.33	28.64	20
Internal medicine	67.60	25.11	150
Family medicine	67.80	27.32	76
Psychiatry	68.06	21.70	21
Radiation oncology	68.65	22.01	16
Otolaryngology	68.80	22.26	32
Ophthalmology	69.11	25.24	30
Anesthesiology	69.59	24.80	59
Emergency medicine	72.94	25.24	81

Table VI
Five lowest rated positive and negative behavioral characteristics

	Rating (1 = not at all, 5 = to a great extent)	SD
Lowest positive behavioral characteristics		
Remains approachable, even when stressed out	3.50	0.81
Stays focused under stress	3.46	0.67
Asks others to do things respectfully	3.42	0.75
Motivates team members to work hard	3.15	0.78
Points out mistakes in a respectful manner	3.14	0.83
Highest negative behavioral characteristics		
Reduced some team memberś job satisfaction	1.65	0.69
Criticizes certain team members behind their back	1.59	0.61
Snaps at others when frustrated	1.56	0.72
Uses offensive gestures when angry	1.55	0.65
Made some team members intimidated or nervous	1.53	0.69

specialists who we examined the baseline LTI scores are lower, suggesting longitudinal, personal approaches, such as tier 3 training, should be undertaken by surgical specialists in order to improve their leadership and teamwork behaviors. Targeted intervention based on individual and specialty differences in behavior shows promise in improving physician-physician and physician-staff interactions.

In the future, we will likely understand the nuances of working with surgical versus medical specialists while they undergo a change process. We found that beyond the differences in baseline LTI scores among physician specialties, there also exists differences in subspecialties. These differences may be owing to further cultural differences among subspecialties beyond a surgical versus medical designation. However, based on our study, it is clear that there is a difference in leadership and teamwork style between the surgical and medical specialties and that these broad categories represent which groups may be more effectively targeted with coaching rather than simply providing a feedback report. As a whole, we need to include sensitivity and awareness that baseline feedback scores may differ depending on specialty, and interventions should be tailored to the individual's specialty and developmental needs.

Working in interprofessional teams is an essential component of modern health care delivery. We have shown that the implementation of a program like PULSE 360 has the potential to improve a critical set of physician skills used for team leadership and communications. The use of MSF in ongoing quality improvement programs directed toward provider collaboration has the potential to drive positive shifts in the culture of safety in medicine.

Limitations

There are several limitations to our study including the lack of a control group, potential selection bias in the assignment of intervention groups, unclear durability of interventions, and lack of granular data on facility type and provider characteristics.

All providers participating in the study were assigned an intervention ranging from a feedback report to individual coaching. In this study, we did not blind any participants to their feedback reports (tier 1 intervention and above), therefore limiting analysis on variation of feedback reports over time without any intervention. However, as there was relatively no improvement or change with just showing people their feedback report (tier 1 intervention), this effectively serves as our control sample.

There is also potential selection bias in that institutions were given the choice to select the tier of intervention. Those with the lowest scores may have had the most potential for actual improvement, whereas those with higher scores had less potential regardless of intervention. However, it is clear that tier 3 interventions improve LTI scores significantly, though the magnitude may be exaggerated because physicians were followed for a maximum duration of only 24 months. It is unclear if ongoing intervention is necessary to effect a consistent change in LTI scores. Future studies would ideally be randomized to minimize both selection bias as well as follow-up after the intervention has ceased to examine durability.

Finally, the lack of granularity in the data with respect to provider characteristics, such as age, years of practice, position within the department, limits our ability to further stratify which providers within subspecialties deserve the most attention. It is also unclear whether there are specific institutional or cultural influences on baseline LTI and effectiveness of interventions. Future studies should include granular data with respect to provider and institutional characteristics.

Physician engagement with the PULSE 360-degree feedback measurement tool is critical to its successful deployment. The current study did not survey physician attitudes toward the evaluation process or the subsequent efforts at education to improve LTI scores. The feedback collection process included administrators and managers, health care provider colleagues, and support staff. Notably, feedback was not collected from patients. Previous studies have demonstrated that patient feedback on communication skills varies based on specialty.[31,32] Future studies may examine how 360-degree feedback correlates with patient feedback. It is also unclear to what extent measuring and attempting to improve patient feedback about physicians using the MSF process might impact relevant changes to patient-centered outcomes. Furthermore, the current study did not measure relevant organizational outcomes related to LTI score improvements, for example patient satisfaction scores or changes in the frequency of the occurrence of safety events or near misses in the practice environment. While the present study clearly shows that low-scoring physicians can improve their feedback scores with intervention, it is unclear to what degree this improvement impacts other relevant leadership and practice outcomes. While the sample in this study shows differences between surgeons and other specialty categories, it cannot be concluded that mean scores for surgeons who did not receive follow-up surveys also would be different. Finally, it is not clear how these outcomes correlate with clinical outcomes such as the National Surgical Quality Improvement Program scores.

Conclusion

In conclusion, modern health care delivery increasingly depends on interdisciplinary care coordination and working in teams. Organizational and cultural transformation in health care therefore requires successfully improving skills and behaviors related to leadership and provider communications. Measuring perceptions regarding physician behavior using the web-based PULSE 360 Program turns a formerly qualitative assessment into a trackable measurement. Baseline MSF scores vary between specialties, with surgical specialists displaying lower LTI scores than primary care and medical specialist colleagues. We have shown that PULSE 360 MSF combined with a tiered system of personalized interventions can improve measurements of communication and leadership over time in physicians who score poorly on the LTI. Surgeons in particular who score poorly show potential to significantly improve scores in response to longitudinal, personalized feedback-based coaching and education. Adding MSF to existing, traditional measurements of physician competency in a comprehensive feedback process may provide physicians with a more meaningful and actionable evaluation process. More research is needed to determine if the LTI scores correlate with clinical outcomes and to determine the durability of the improvement provided by these behavioral innovations.

Conflict of interest/Disclosures

Drs. Gregory and Harmon are employees of the Physicians Development Program, sole proprietor of the PULSE 360 Survey and the PULSE 360 Program.

Although some of our authors are employed by Pulse 360, the data analysis method was determined by only University faculty and students and the analysis was completed in its entirety at the University of Buffalo. All raw data was made available to the University personnel who designed and performed the analysis.

Funding/Support

None.

Acknowledgments

Hu and Mullin had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

References

1. Kozlowski SW, Ilgen DR. Enhancing the effectiveness of work groups and teams. *Psychol Sci Public Interest*. 2006;7:77–124.
2. Manser T. Teamwork and patient safety in dynamic domains of healthcare: a review of the literature. *Acta Anaesthesiol Scand*. 2009;53:143–151.
3. Weller J, Boyd M, Cumin D. Teams, tribes and patient safety: overcoming barriers to effective teamwork in healthcare. *Postgrad Med J*. 2014;90:149–154.
4. Weaver SJ, Dy SM, Rosen MA. Team-training in healthcare: a narrative synthesis of the literature. *BMJ Qual Saf*. 2014;23:359–372.
5. Baggs JG, Ryan SA, Phelps CE, Richeson JF, Johnson JE. The association between interdisciplinary collaboration and patient outcomes in a medical intensive care unit. *Heart Lung*. 1992;21:18–24.
6. Young GJ, Charns MP, Desai K, et al. Patterns of coordination and clinical outcomes: a study of surgical services. *Health Serv Res*. 1998;33:1211–1236.
7. Cooper WO, Guillamondegui O, Hines OJ, et al. Use of unsolicited patient observations to identify surgeons with increased risk for postoperative complications. *JAMA Surg*. 2017;152:522–529.
8. Rabøl LI, Andersen ML, Østergaard D, Bjørn B, Lilja B, Mogensen T. Descriptions of verbal communication errors between staff. An analysis of 84 root cause analysis-reports from Danish hospitals. *BMJ Qual Saf*. 2011;20: 268–274.
9. Singh H, Thomas EJ, Petersen LA, Studdert DM. Medical errors involving trainees: a study of closed malpractice claims from 5 insurers. *Arch Intern Med*. 2007;167:2030–2036.
10. Holmboe ES, Edgar L, Hamstra S. The milestones guidebook, 2016. http://www.acgme.org/Portals/0/MilestonesGuidebook.pdf. Accessed January 14, 2019.
11. Brutus S, Fleenor JW, London M. Does 360-degree feedback work in different industries?: A between-industry comparison of the reliability and validity of multi-source performance ratings. *J Manag Dev*. 1998;17:177–190.
12. Drew G. A “360” degree view for individual leadership development. *J Manag Dev*. 2009;28:581–592.
13. Donnon T, Al Ansari A, Al Alawi S, Violato C. The reliability, validity, and feasibility of multisource feedback physician assessment: a systematic review. *Acad Med*. 2014;89:511–516.
14. Gregory PJ, Ring D, Rubash H, Harmon L. Use of 360° feedback to develop physician leaders in orthopaedic surgery. *J Surg Orthop Adv*. 2018;27: 85–91.
15. Stevens S, Read J, Baines R, Chatterjee A, Archer J. Validation of multisource feedback in assessing medical performance: a systematic review. *J Contin Educ Health Prof*. 2018;38:262–268.
16. Brinkman WB, Geraghty SR, Lanphear BP, et al. Effect of multisource feedback on resident communication skills and professionalism: a randomized controlled trial. *Arch Pediatr Adolesc Med*. 2007;161:44–49.
17. Hageman MG, Ring DC, Gregory PJ, Rubash HE, Harmon L. Do 360-degree feedback survey results relate to patient satisfaction measures? *Clin Orthop Relat Res*. 2015;473:1590–1597.
18. Lagoo J, Berry WR, Miller K, et al. Multisource evaluation of surgeon behavior is associated with malpractice claims. *Ann Surg*. 2019;270:84–90.
19. Council of Academic Hospitals of Ontario (CAHO). 360-degree physician performance review toolkit. 2009. http://caho-hospitals.com/wp-content/uploads/2014/02/CAHO-360-Degree-Physician-PerformToolkit2009.pdf. Accessed December 3, 2019.
20. Nurudeen SM, Kwakye G, Berry WR, et al. Can 360-degree reviews help surgeons? Evaluation of multisource feedback for surgeons in a multi-institutional quality improvement project. *J Am Coll Surg*. 2015;221:837–844.

21. Weigelt JA, Brasel KJ, Bragg D, Simpson D. The 360-degree evaluation: increased work with little return? *Curr Surg*. 2004;61:616–626; discussion 627-628.
22. Lockyer J. Multisource feedback in the assessment of physician competencies. *J Contin Educ Health Prof*. 2003;23:4–12.
23. Karsten MA. Coaching: an effective leadership intervention. *Nurs Clin North Am*. 2010;45:39–48.
24. Tumerman M, Carlson LM. Increasing medical team cohesion and leadership behaviors using a 360-degree evaluation process. *WMJ*. 2012;111:33–37.
25. Yama BA, Hodgins M, Boydell K, Schwartz SB. A qualitative exploration: questioning multisource feedback in residency education. *BMC Med Educ*. 2018;18:170.
26. Wang L, Zhang Z, McArdle JJ, Salthouse TA. Investigating ceiling effects in longitudinal data analysis. *Multivariate Behav Res*. 2008;43:476–496.
27. Freedman DA. On the so-called "Huber sandwich estimator" and "robust standard errors." *Am Stat*. 2006;60:299–302.
28. Thach EC. The impact of executive coaching and 360 feedback on leadership effectiveness. *Leader Organ Dev J*. 2002;23:205–214.
29. Cooper JB. Critical role of the surgeon-anesthesiologist relationship for patient safety. *Anesthesiology*. 2018;129:402–405.
30. Villafranca A, Hamlin C, Enns S, Jacobsohn E. Disruptive behaviour in the perioperative setting: a contemporary review. *Can J Anaesth*. 2017;64:128–140.
31. Bindels E, Boerebach B, Van der Meulen M, et al. A new multisource feedback tool for evaluating the performance of specialty-specific physician groups: validity of the group monitor instrument. *J Contin Educ Health Prof*. 2019;39:168–177.
32. Quigley DD, Elliott MN, Farley DO, Burkhart Q, Skootsky SA, Hays RD. Specialties differ in which aspects of doctor communication predict overall physician ratings. *J Gen Intern Med*. 2014;29:447–454.