Improving the Reliability and Validity of Accreditation and Certification Scoring

June 11, 2020

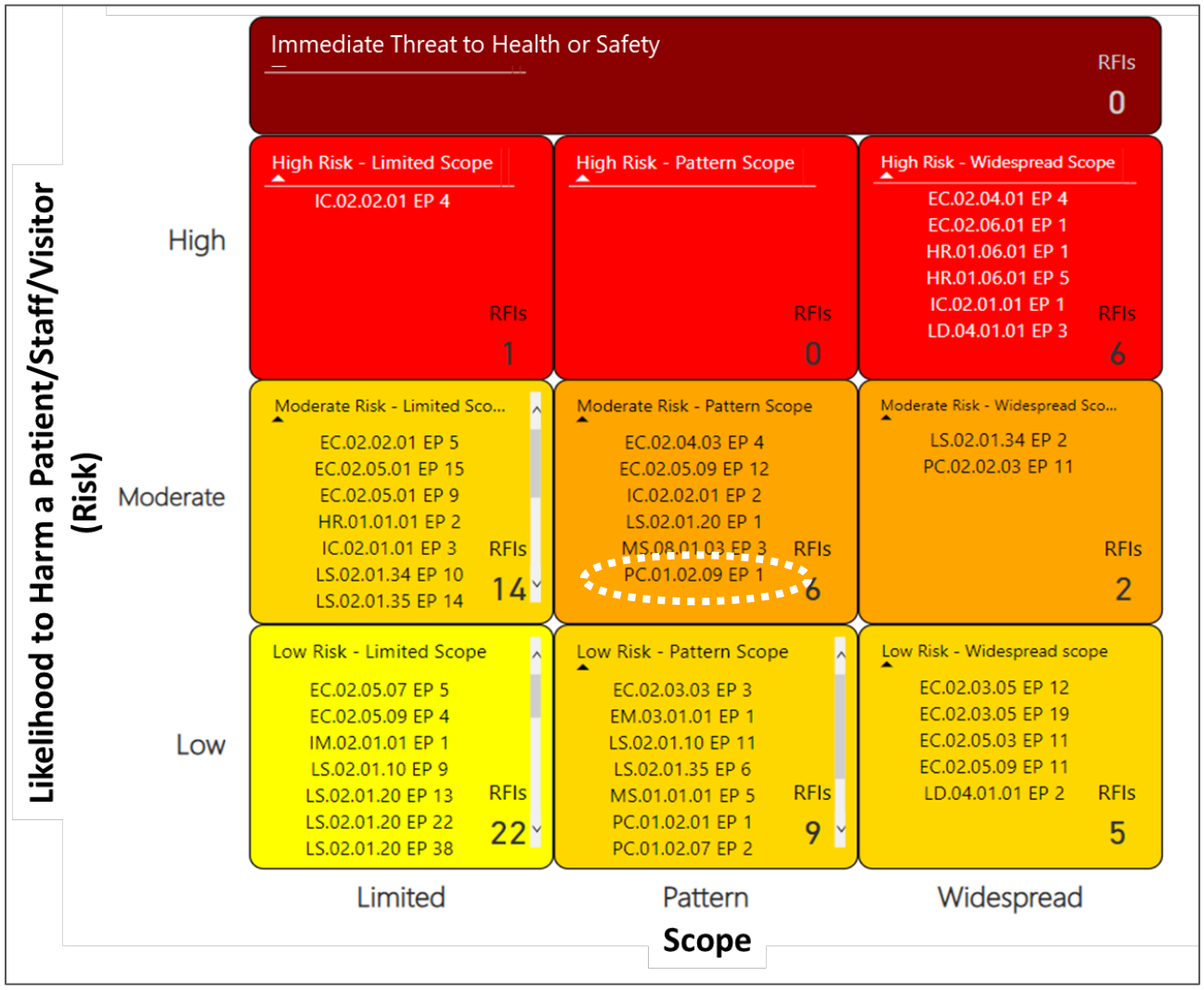
Consistency is one of the hallmarks of a highly reliable organization, regardless of whether that organization is in aviation, manufacturing, nuclear power, or healthcare. Policies and protocols must be developed, rigorous training and competency testing must be conducted, and monitoring systems must be implemented to ensure that processes occur as anticipated and the desired outcomes are achieved. The Joint Commission believes all healthcare organizations should work to become more highly reliable.

But what about our own accreditation survey and certification review processes? We have always had well-established policies and protocols, and we rigorously train surveyors, including classroom time, online learning, real-world experience and mentorship on surveys. (Note that for brevity, we use the accreditation term 'surveyor' and 'survey' throughout this article; these terms should be interpreted to include 'reviewer' and 'review', the terms associated with certification"). However, until recently, we had limited ability to track variations in surveyor scoring, which hampered our efforts to improve consistency. Would different surveyors spend roughly the same amount of time reviewing different areas of an organization? Were some surveyors "tougher" than others and score deficiencies more aggressively? We had no good way of knowing.

When The Joint Commission developed the Survey Analysis for Evaluating Risk (SAFER) matrix™ in 2017, the need to develop ways to ensure consistency became even more important. This new scoring method goes beyond simply identifying a problem with standards compliance and allows surveyors to classify the scope (i.e., how widespread the problem is) and risk (i.e., how likely it is to cause harm) of the deficiency.

The SAFER matrix is designed to provide a more relevant assessment of the risk a survey observation carries. Sometimes, the description of an issue gives the spurious impression that it is relatively minor. However, the problem can lead to significant risks and serious implications for patients. The SAFER matrix™, provides a mechanism to help organizations prioritize resources and focus corrective actions on areas that could have the most significant impact on patients (**Figure 1**).

Figure 1: Example of SAFER Matrix – Placement of Survey Findings and Link to Complete Citation



PC.01.02.09 EP 1: The organization assesses the patient who may be a victim of possible abuse and neglect. The organization uses criteria to identify those patients who may be victims of physical assault, sexual assault, sexual molestation, domestic abuse, or elder or child abuse and neglect. Note: Criteria can be based on age, sex, and circumstance.

Observation: During tracer activities in the Pediatric Acute Care Unit, it was observed that the record of an outpatient who had undergone an open reduction internal fixation of a wrist fracture did not contain documentation of an abuse or neglect assessment. In conversation with staff during tracer activity in the PACU, OR and inpatient floor throughout the day, the staff caring for scheduled outpatients who are discharged home or subsequently admitted, do not routinely screen for abuse and neglect. Patients admitted to inpatient status from the Emergency Department direct to floor were screened by inpatient nurses per conversation with the inpatient nursing staff. RISK

While the concept of the SAFER Matrix was well received by customers, the need to place all survey findings within a 10-box matrix heightened concerns about inconsistency. For this reason, as the SAFER Matrix was tested and rolled out, The Joint Commission simultaneously implemented a new quality improvement program to address the fundamental challenge of scoring consistency, as well as the new challenges created by the introduction of the SAFER matrix. This paper describes the multiple strategies we

implemented over the last three years to improve scoring consistency and validity, and we summarize the major improvements that have resulted from these efforts. We believe that this approach is a model that all accrediting organizations should follow.

Methods

Measuring Variation in SAFER Scoring

In order to compare surveyor scoring patterns, survey findings from all full-survey events over a 12-month period are aggregated (as counts) based upon their assignment to the 10-box SAFER matrix. A Chi-Square test is then used to compare the distribution of counts for each individual surveyor to the distribution of counts for their peer group. Each peer group is based upon the specific accreditation program (e.g., Hospital, Home Health Care, Behavioral Health Care) and the surveyor role during the survey process (e.g., engineer, physician, nurse). So, the performance of an individual surveyor, who completed 25 full surveys within the ambulatory health care program as an engineer would be compared to the distribution of findings for other engineers who were surveying ambulatory health care organizations during the same time period. We refer to the Chi-Square calculation as the *Variation Index*. It provides a numeric value for each individual surveyor (or more than one, if they surveyed multiple programs) where a score of zero (0) indicates perfect alignment with the peer group and increasing values indicate greater variation from the peer group.

Given that surveyors are evaluating different healthcare organizations, it is assumed that all surveyors will differ from the peer group norm to some degree. Therefore, in order to aid in the interpretation of the Variation Index, and to prioritize supervision activities, the surveyor Variation Index values are converted into Z-Scores. A Z-Score of zero (0) indicates that a surveyor's variation is at the average for their peer group. Negative scores indicate that the surveyor's variation is lower than average, whereas positive scores indicate that variation is higher than average. For example, a positive Z-score of 1.5 indicates the surveyor's variation that is 1.5 standard deviations above the average for their peer group.

To identify specific patterns of variation that can be used to guide discussions with individual surveyors, SAFER scoring is further broken down along each dimension of the SAFER matrix: "Scope" (Limited, Pattern Widespread or Immediate Threat to Health or Safety [ITHS]) and "Likelihood to Harm" (Low, Moderate, High or ITHS). Variation indices and Z-scores are calculated for both Scope and Risk, in addition to the index for overall SAFER variation.

Measuring Variation in Survey Domain Scoring

In addition to evaluating consistency on the SAFER Matrix, we also assess individual variation in survey domain scoring patterns. During the survey process, accreditation requirements are classified by various quality and safety domains (e.g., infection control, leadership, life safety code, medication management, patient care, etc.). To assess the degree to which surveyors and survey teams are consistently evaluating these domains during survey events, a similar process for calculating a Variation Index is applied. In this case, the scoring patterns of individual surveyors (i.e., counts of survey findings within each quality and safety domain) are compared to the scoring patterns of their peer group for each accreditation program. Using this approach, it is possible to determine if an individual surveyor has a tendency to focus more intensively on certain topic areas and/or less intensively on others. On the Domain Scoring dimension, higher Z-scores indicate a greater degree of variation in the proportion of findings across domains, relative to other surveyors in the same role and same accreditation program.

Surveyor Coaching Tool

To understand and address scoring inconsistencies, a data visualization tool was developed to give supervisors real-time data that can be used to illustrate performance differences and assist with coaching efforts. The application was developed and pilot tested in 2017. Throughout the development and

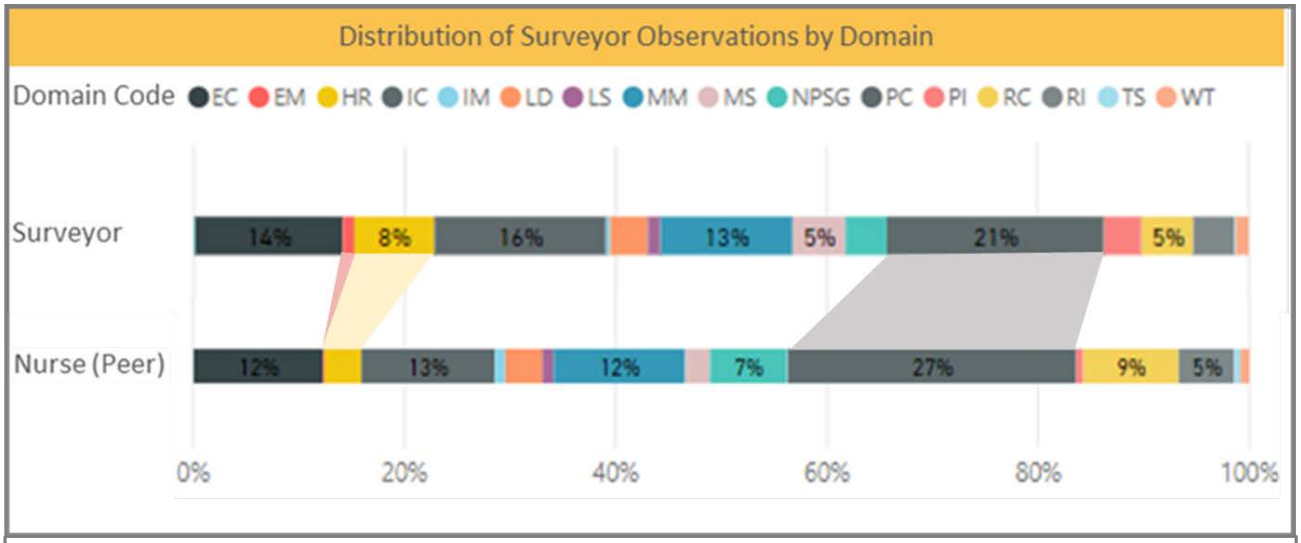
testing process, the data visualization team met regularly with the supervisory staff of the Joint Commission's surveyors. Feedback from these sessions (e.g., questions about data, requests for additional details, concerns about interpretation) was used to refine the application. By the end of 2017, the application referred to as the *Surveyor Coaching Tool* evolved into a three-step process for visualization and coaching.

1. Step one: Provides supervisors with an annotated list of the surveyors using colored flags to draw attention to individual surveyors who are variation outliers in their scoring patterns.
2. Step two: Direct visual comparison of scoring patterns for an individual surveyor. The visual presentation allows supervisors to assess the magnitude of differences when comparing an individual surveyor to their peer group. **Figure 2** illustrates this process for Domain Variation. Note that in this example, the surveyor tends to score leadership (LD) standards more frequently than peers and scores infection control (IC) and patient care (PC) less frequently than peers. The same process is used to draw comparisons in SAFER Matrix scoring (Risk and Scope).
3. Step three: Supervisors use the interactive features of the application to drill directly into the individual findings that led to the variation in scoring. Supervisors begin by comparing individual scoring patterns from surveyor to surveyor (See **Figure 3**). In the example, SAFER placement of findings for the same infection control (IC) standard is compared among surveyors. By clicking on the chart, supervisors can see the actual survey findings rather than just the numbers, this allows them to conduct a direct inspection and draw conclusions about the cause of the variation (e.g., misinterpretation of a standard or SAFER placement).

Interventions to Reduce Inconsistency

As the Surveyor Coaching Tool was formally introduced at the beginning of 2018, supervisors were instructed to focus on outliers (i.e., surveyors with Z-scores of 3 or higher). The coaching tool was their mechanism to visually identify outliers, rapidly determine the specific point of deviation for a surveyor, and drill into their individual findings. The ability to dig into the individual citations was an essential feature for a number of reasons. First, it allowed supervisors to engage their surveyors in discussions about what they were seeing and scoring, rather than focusing on statistics. Second, since SAFER scoring was new, supervisors were encouraged to keep an open mind about the nature of scoring discrepancies. It was assumed that in many cases, scoring variation may be justified (i.e., it could be attributed to unique circumstances associated with a survey event). In other situations, discussions might lead to real-time education with a specific surveyor, or to the clarification of definitions that benefited the entire peer group. In fact, the coaching process often took place while supervisors and supervisees shared computer screens over the phone so that the visual comparison of surveyor performance and the drilling into specific observations was collaborative (like coaches and athletes studying game film together). This supervision practice tended to promote greater transparency and enhance trust. As supervisors gained experience with the tool, the outlier threshold was reduced to a Z-score of 2.5.

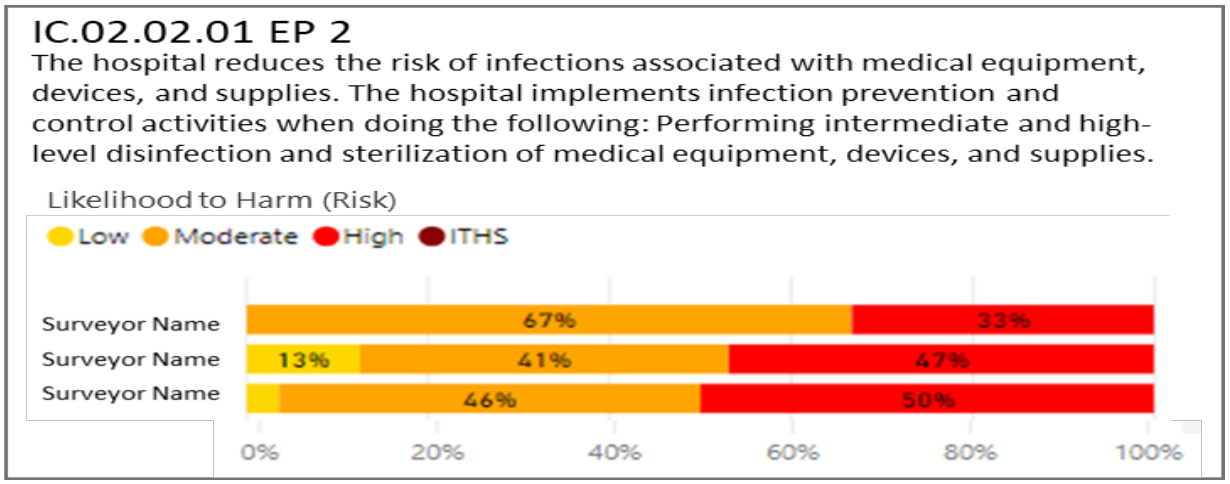
Figure 2: Data Visualization Tool for Supervisors to Compare Individual Surveyors to Their Peers



EC=Environment of Care; EM=Emergency Management; HR=Human Resources; IC=Infection Control; IM=Information Management; LD=Leadership; LS=Life Safety; MM=Medication Management; MS=Medical Staff; NPSG=National Patient Safety Goal; PC=Patient Care; PI=Performance Improvement; RC=Record of Care; RI=Patient Rights; TS=Transplant Services; WT=Waived Testing

When an individual surveyor's variation has been flagged as significantly different from that of their peer group, supervisors are able to visually inspect those differences. In the chart above, an individual surveyor appears less likely to score issues related to Infection Control (IC) or Patient Care (PC) and more likely to identify findings associated with Leadership (LD) when compared to other nurses conducting surveys for the same accreditation program. Supervisors can then compare the number of findings issued by a surveyor and drill directly into the data to see actual survey findings, in order to draw conclusions about the cause of the variation observed.

Figure 3: Data Visualization Tool for Comparing Individual Surveyor SAFER Scoring Patterns on a Single Standard and Element of Performance



By focusing on a single standard and element of performance, supervisors can compare scoring of individual surveyors for a similar topic area. In the chart above, the first surveyor appears to be less likely to score this issue as high risk, when compared to the other two surveyors on the list. Supervisors can then drill into the individual written findings to determine if scoring was accurate or if additional education is needed.

By mid-2019 the application was further enhanced so that supervisors could track the impact of their supervision efforts by monitoring changes in surveyor variation over time (e.g., year-over-year; see **Figure 4**). And by the end of 2019, supervisors were using the coaching tool to assist with annual evaluations.

Figure 4: Change in Individual Surveyor Variation Scores Over Time

Name	Program	Year	Survey Count	Findings	SAFER Z-Score	Risk Z-Score	Scope Z-Score	Domain Z-Score
Surveyor 1	Hospital	2018	46	1226	2.22	2.89	1.95	1.29
Surveyor 1	Hospital	2019	39	948	0.06	0.21	0.20	0.52
Surveyor 1	Hospital	2020	5	107	-0.34	-0.55	-0.08	-0.23

Note: The data points in the line charts above display an individual surveyor's Z-Scores for their Variation Indices (SAFER Overall, SAFER Risk, SAFER Scope and Survey Domains. In the data table below the charts, the orange background in one cell identifies a Variation Index score that was greater than 2.5 standard deviations from average variation for the surveyor's peer group (e.g., 2018 SAFER Risk Z-Score). This can also be observed in the data point in the SAFER Risk line chart that crosses the 2.5 Standard Deviation threshold.

According to supervisors, efforts to address scoring inconsistencies tended to require different solutions, depending upon the nature of the underlying problem. These generally fell into one of three categories:

1. Individual Coaching and Education: In the simplest cases a surveyor may have misinterpreted scoring guidelines, and brief education was sufficient to clarify and correct a problem.
2. Group Education: In some cases, however, inconsistency between a

surveyor and their peer group helped to identify a need for better education across the peer group. For example, one surveyor was identified as an outlier for scoring a higher proportion of findings as "widespread" in scope. After reviewing the written findings and discussing this with the surveyor, it became clear that his interpretation was correct and the peer group needed education. Specifically, when a health care organization was missing a key element from a required policy, many surveyors in the peer group were inclined to score this problem as limited in scope (since it was a single element of a single policy). In contrast, this kind of compliance issue should have been scored as "widespread", since a policy has the potential to impact the entire organization and/or a large number of patients or staff.

3. Improving the Clarity of Standards and/or Scoring Guidance: The most challenging cases tended involve situations in which context could play a major role in SAFER scoring. For example, in several instances a nearly identical finding could be interpreted to be of low, moderate or high risk. In one situation, the absence of a battery powered light fixture near an emergency generator could be scored as low risk – if that generator was part of a larger suite of generators that shared power loads. Alternatively, it could be considered high risk if the generator was the single power source for a critical area service area, and it was located in an interior room without an ambient light source.

In response to these types of issues, The Joint Commission began to acquire a growing library of challenging scoring examples for standards that exhibited higher degrees of inconsistency (i.e., those judged to need more education or clarification). In many cases, the examples were reviewed by multiple experts who provided a rationale for SAFER placement recommendations. Over time, collections of these examples were incorporated into educational efforts as tools that could be used to help to guide placement in the SAFER Matrix. During 2019, many of these examples were also being routinely added to the electronic scoring system that surveyors use during the survey process. For example, the National Patient Safety Goal related to Suicide Prevention (NPSG 15.01.01) requires that organizations to follow written policies and procedures addressing the care of patients at risk for suicide. At a minimum, these policies should include guidelines for re-assessment and procedures monitoring patients who are at risk for suicide, and requirements for training and competency assessment of staff who care for patients at risk for suicide (Element of Performance #5). During the survey process, as a surveyor begins to record a compliance issue related to this requirement, he/she is presented with several examples portraying varying levels of risk. The failure to reassess a high risk suicide patient in accordance with the policy, or a situation in which a 1:1 monitor leaves a high risk patient unattended in the bathroom would be examples of "High" risk findings on the SAFER Matrix. In contrast, a situation in which nurses responsible for completing suicide risk reassessments did not have evidence of training or competency assessment would be categorized as "Moderate" risk.

These examples, which are frequently developed with input from experts, often include rationale statements to explain how the situational context and other factors justified placement at a specific risk level. The rationale statements are an important component, since it is not possible to create examples that cover every possible situation encountered by the survey team. Therefore, the rationale statements help to guide the systematic thought process needed to assess risk.

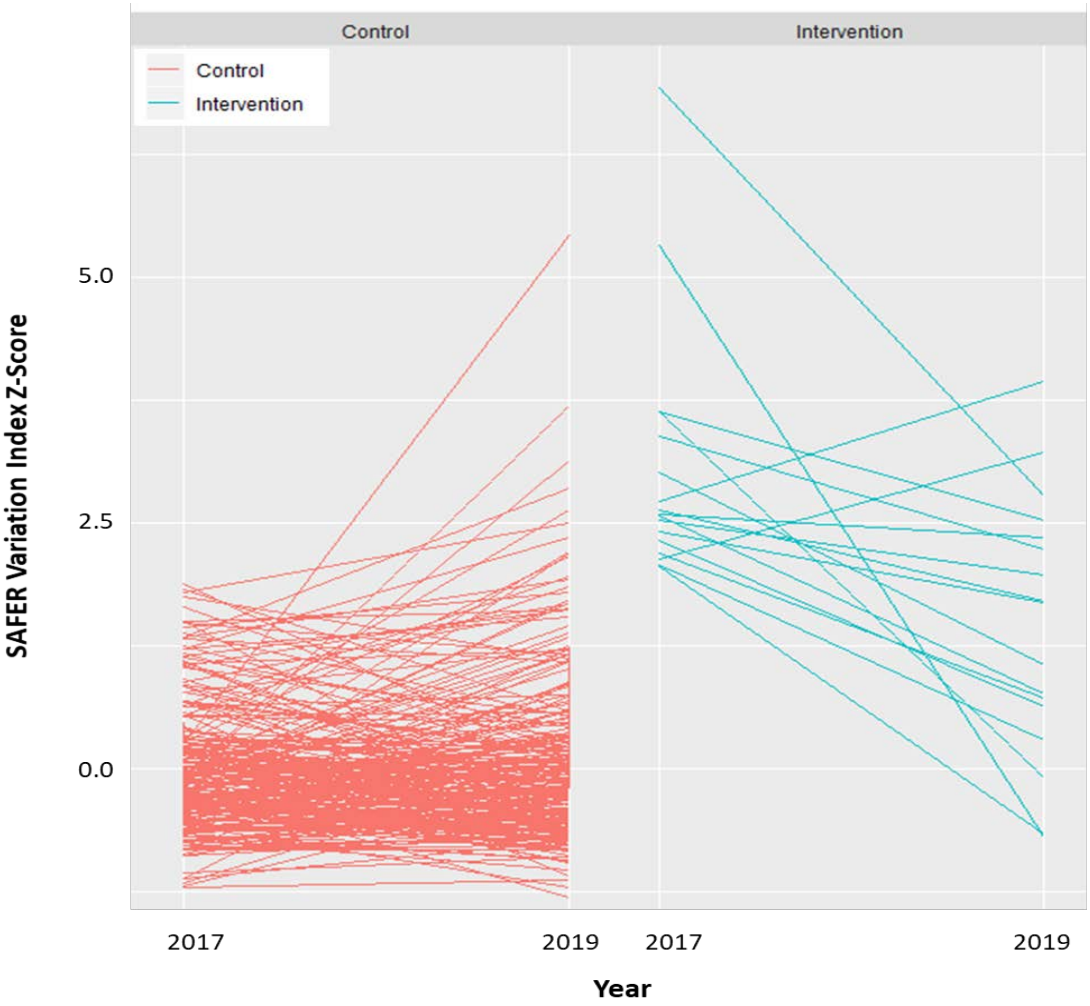
Results

The SAFER scoring system was introduced in 2017, and by 2018 enough data had been collected to begin calculating the Variation Indices for most surveyors. By mid-2018, supervisors were encouraged to use these data in regular supervision activities by using the Surveyor Coaching Tool. This allowed us to

assess the impact of these multi-faceted efforts on surveyor consistency.

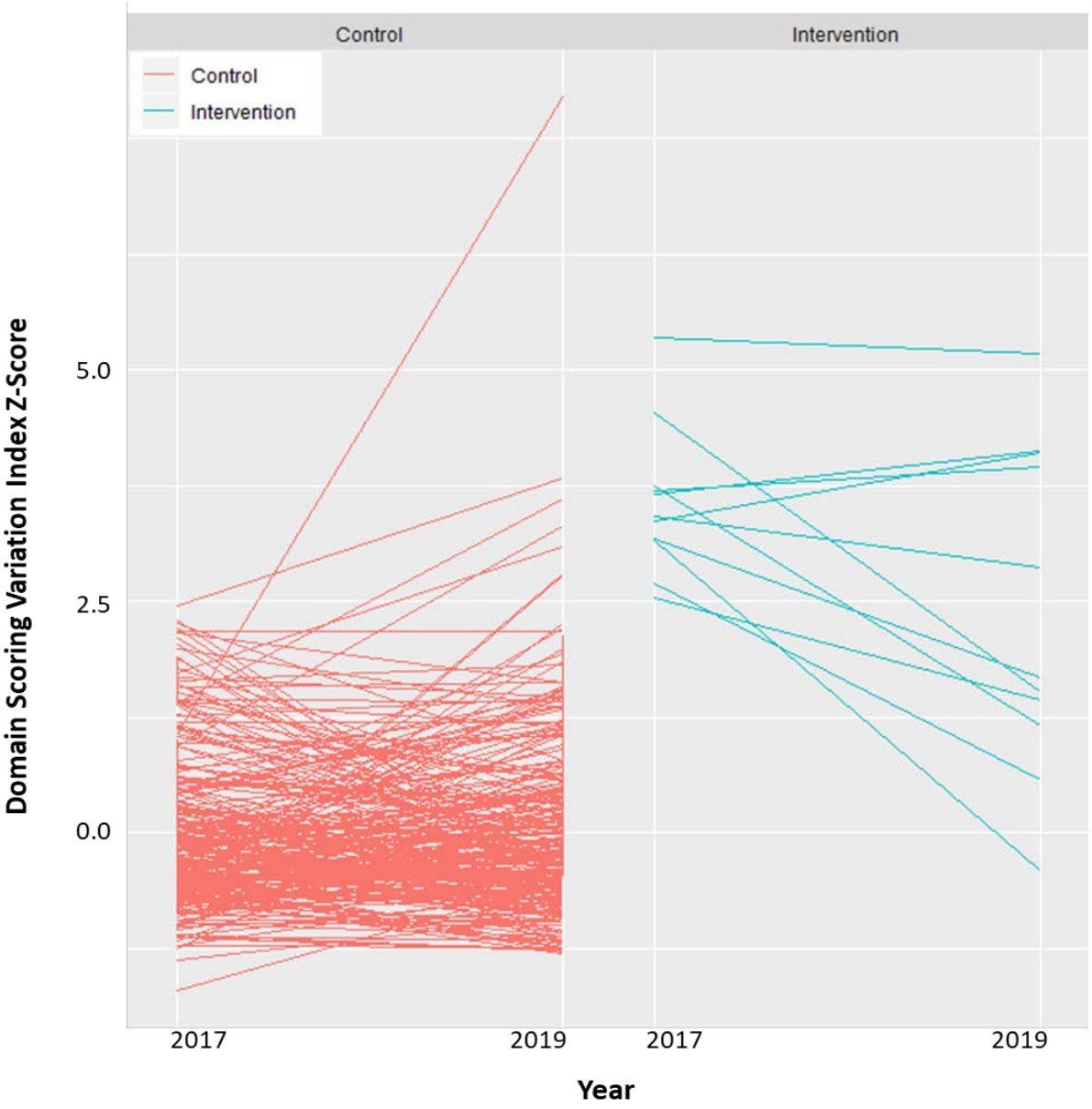
In order to quantitatively assess changes in surveyor consistency, surveyor Z-scores were compared between 2017 and 2019. To be included in the analysis, surveyors must have performed a minimum of five full survey events in both time periods. Surveyors with fewer than five full surveys in either year, or those that could not be matched in the two time periods, were excluded from the analysis. Surveyors were then grouped into an intervention group (i.e., those surveyors identified as outliers in 2017, who became the focus of coaching and education efforts) and a control group (surveyors not identified as outliers in 2017, who received supervision as usual). Paired t-tests were used to compare the differences in the 2017 and 2019 Z-scores. The comparison for SAFER Scoring revealed that surveyors in the intervention group exhibited statistically significant improvement as compared to controls (t-test = 4.77, df = 20.72, p-value < 0.001). See Figure 5 for a graphic depiction of these differences.

Figure 5: Changes in Overall SAFER Variation from 2017-19 for Surveyors Receiving Coaching Intervention



This analytic approach was repeated to evaluate improvement in Domain scoring. Surveyors in the intervention group improved significantly when compared controls (t-test = 3.94, df = 15.32, p-value = 0.001). See **Figure 6** for a graphic depiction of these differences.

Figure 6: Changes in Domain Scoring Variation from 2017 to 2019 for Surveyors Receiving Coaching Intervention



It is worth noting that, while coaching and education efforts appeared to have a positive impact on the surveyors identified as outliers in 2017, new outliers can be readily identified in the 2019 data. Achieving and maintaining high reliability is an ongoing process, not a one-time project. For this reason, new outliers are flagged on a quarterly basis so that supervisors may continue to re-direct coaching efforts as new potential problems are identified.

Finally, to compare the impact of improvement efforts at the accreditation program level, an effect size was calculated to assess change in the average SAFER Variation Index for all surveyors within the program area (matched by year, for surveyors with a minimum of 5 surveys during the year). Effect size was calculated for matched pairs using a repeated measures correction (ES_{RMC}) that is derived from a t-test of two correlated (paired) means corrected for the covariation of those means. [i] The approach

was selected because it provides a stable and moderate estimate of effect size when compared to many "raw" calculations of effect size.[ii] Results are displayed in **Table 1.**

Table 1: SAFER Scoring: Effect Size of Change in Variation Index 2017 - 2019

Program		2017	2019	ES	CI
Hospital	Mean	95.3	89.3	-0.04	0.18
n=164	Std Dev	164.3	122.9		
Behavioral	Mean	118.6	106.0	-0.13	0.13
n=33	Std Dev	105.3	87.2		
Ambulatory	Mean	75.5	63.7	-0.16	0.18
n=53	Std Dev	80.6	60.7		
Nursing Care	Mean	70.7	119.3	0.72	0.09
n=8	Std Dev	39.1	75.2		
Laboratory	Mean	151.3	126.4	-0.16	0.10
n=19	Std Dev	143.9	156.2		
Home Care	Mean	84.6	78.5	-0.09	0.19
n=70	Std Dev	76.7	62.9		

n = number of surveyors in analysis; ES = Effect Size; CI = 95% Confidence Interval

Negative effect sizes represent improvement between the time periods, and improvement was observed for all accreditation programs except the Nursing Care program (which is difficult to interpret due to the small number of surveyors in the cohort). While observed effects have been modest to date, interventions were initially focused on a small number of surveyors (outliers) and education efforts implemented in 2019 are likely to have more noticeable impacts in subsequent years.

Conclusions

Performance measurement with feedback (audit and feedback), including benchmarking compared to peers, is a central strategy for quality improvement around the world. Just as The Joint Commission uses this methodology with our ORYX Measurement System and other measurement programs to help our accredited and certified organizations improve, we undertook this effort to do the same with our surveyors. Specifically, we needed to ensure that surveyors have a consistent approach to how they assess the domains of safety for our standards and how they rate the likelihood to harm and scope of findings using the SAFER matrix. The tools we developed to identify surveyors with scoring patterns substantially different than their peers and provide them with concrete examples of improvement opportunities led to improved consistency for most of the surveyors who were identified as outliers.

The measurement system was not designed to systematically measure the validity or accuracy of surveyor scoring. However, although our main focus was on consistency, the tools also allowed supervisors to assess the validity of surveyors' scoring more easily; this sometimes identified surveyors whose scoring was rational and justified, and the examples from these surveyors were then used to set a standard for others to follow. In addition, we developed prototypes of commonly encountered deficiencies found on survey and recommended SAFER scoring levels for likelihood to harm and scope.

We believe that other accrediting/certifying organizations are likely to have similar variations in scoring patterns across surveyors as we found at baseline, especially organizations that are trying to implement systems similar to

the SAFER matrix to classify the risk and scope of findings. Training alone is not enough, and direct observation by supervisors to provide feedback cannot substitute for comprehensive assessment of scoring patterns. Just as The Joint Commission is encouraging health care organizations to work towards high reliability, accrediting organizations need to do the same by routinely measuring surveyor variations and continuously working to improve.

Questions? Contact Scott Williams, PsyD
swilliams@jointcommission.org

References

[i] Dunlap WP, Cortina JM, Vaslow JB and Burke MJ. Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods*. 1996; 1(2):170-177.

[ii] Seidel JA, Miller SD and Chow DL. Effect size calculations for the clinician: Methods and comparability. *Psychotherapy Research*. 2014;24(4):470-84.