

Advancing Measurement of Patient Safety Culture

*Liane Ginsburg, Debra Gilin, Deborah Tregunno,
Peter G. Norton, Ward Flemons, and Mark Fleming*

Objective. To examine the psychometric and unit of analysis/strength of culture issues in patient safety culture (PSC) measurement.

Data Source. Two cross-sectional surveys of health care staff in 10 Canadian health care organizations totaling 11,586 respondents.

Study Design. A cross-validation study of a measure of PSC using survey data gathered using the Modified Stanford PSC survey (MSI-2005 and MSI-2006); a within-group agreement analysis of MSI-2006 data.

Extraction Methods. Exploratory factor analyses (EFA) of the MSI-05 survey data and confirmatory factor analysis (CFA) of the MSI-06 survey data; Rwg coefficients of homogeneity were calculated for 37 units and six organizations in the MSI-06 data set to examine within-group agreement.

Principal Findings. The CFA did not yield acceptable levels of fit. EFA and reliability analysis of MSI-06 data suggest two reliable dimensions of PSC: *Organization leadership for safety* ($\alpha = 0.88$) and *Unit leadership for safety* ($\alpha = 0.81$). Within-group agreement analysis shows stronger within-unit agreement than within-organization agreement on assessed PSC dimensions.

Conclusions. The field of PSC measurement has not been able to meet strict requirements for sound measurement using conventional approaches of CFA. Additional work is needed to identify and soundly measure key dimensions of PSC. The field would also benefit from further attention to strength of culture/unit of analysis issues.

Key Words. Patient safety culture measurement, patient safety climate measurement, culture strength, modified Stanford safety culture survey

The importance of patient safety culture (PSC) measurement in health care is well documented (Pronovost and Sexton 2005; Gaba, Singer, and Rosen 2007; Pace 2007) and measures of PSC are proliferating (Hutchinson et al. 2006). In part, growth in this area parallels increasing external pressure on health care organizations from accreditation and other safety agencies (Pronovost et al. 2006) in the United States (Joint Commission on Accreditation of Healthcare Organizations [JCAHO], Agency for Healthcare

Research and Quality [AHRQ]), the United Kingdom (the National Health Service [NHS]) and Canada (Canadian Centre for Health and Safety in Agriculture [CCHSA]). The promise of benchmarking is also attracting organizations to measure PSC. Various tools exist to measure PSC/climate¹ (Nieva and Sorra 2003; Ginsburg et al. 2005; Sexton et al. 2006; Singer et al. 2007). In industries outside of health care, safety culture measurement has, on the whole, been developed and used in a research context, either looking at models and measures (Guldenmund 2000, 2007; Cooper and Phillips 2004) or testing relationships between safety culture and other variables (Zohar 2000, 2002; Naveh, Katz-Navon, and Stern 2005; Zohar and Luria 2005; Katz-Navon, Naveh, and Stern 2006; Zohar et al. 2007). Accordingly, in these settings, attention has been paid to methodological matters such as factor structure and unit of analysis issues. In health care, the focus of PSC measurement literature has largely been on system improvement. In the current paper, we draw on the broader organizational literature as it helps identify opportunities for advancing PSC measurement both in terms of (1) psychometric rigor and (2) important unit of analysis and strength of culture issues.

As a construct, safety culture has been defined in a variety of ways in health care and other industries. Some see safety culture as patterns of responses to problems (Westrum 2004) while others define it more narrowly, focusing on the key dimensions of unit and organizational leadership's prioritization of safety (Zohar 2000). PSC is sometimes broadly conceptualized to include sub-dimensions such as learning, reporting, and blame orientation (Reason 1997; Cooper 2000; Hofmann and Mark 2006). Sometimes, the definitions also include more distant dimensions such as job satisfaction (Sexton et al. 2006) and staffing (Nieva and Sorra 2003). As a result of these differing concepts of safety culture it has been defined and measured in numerous ways in health care (Colla et al. 2005; Fleming 2005; Flin et al. 2006;

Address correspondence to Liane R. Ginsburg, Ph.D., Associate Professor, School of Health Policy and Management, Faculty of Health, York University, HNES Building 413, 4700 Keele Street, Toronto, ON, Canada M3J 1P3; e-mail: lgins@yorku.ca. Debra Gilin, Ph.D., Assistant Professor, and Mark Fleming, Ph.D., Associate Professor, are with the Department of Psychology, Saint Mary's University, Halifax, NS, Canada. Deborah Tregunno, R.N., Ph.D., Assistant Professor, is with the School of Nursing, Faculty of Health, York University, Toronto, ON, Canada. Peter G. Norton, M.D., C.C.F.P., F.C.F.P., Professor, is with the Department of Family Medicine, Faculty of Medicine, Calgary, AB, Canada. Ward Flemons, M.D., F.R.C.P.C., F.A.C.P., Vice-President, is with the Quality Safety and Health Information, Calgary Health Region and is also Clinical Professor of Medicine, Faculty of Medicine, University of Calgary, Calgary, AB, Canada.

Sexton et al. 2006) and the broader safety literature (e.g., Zohar 2000; Hofmann and Mark 2006; Guldenmund 2007).

Researchers have sought to advance measurement of PSC in health care and establish construct validity either through the use of relational approaches that focus on convergent and discriminant validity (Singer et al. 2007) or through efforts to identify stable, psychometrically sound measures using exploratory and confirmatory factor analysis (EFA and CFA) (Sorra and Nieva 2003; Naveh et al. 2005; Sexton et al. 2006). While these studies represent advancements in the field of PSC measurement, strong evidence of psychometric rigor has not yet been published for PSC measurement (Flin 2007) or broader organizational safety culture measurement (Guldenmund 2007) according to traditionally accepted psychometric standards that require the use of independent samples for cross validation (i.e., EFA and CFA must be performed on separate samples) (Hu and Bentler 1999). There are at least two perspectives that may explain some of the difficulties researchers have faced trying to confirm a stable set of PSC factors (dimensions). Coyle, Sleeman, and Adams (1995) raise the question of whether a safety culture factor structure may be population specific (e.g., profession specific or even organization/setting specific). Alternatively, inability to confirm a stable PSC factor structure may be due to the fact that PSC is ill-defined as a construct—as James Reason often states, safety culture “has the definitional precision of a cloud” (1997). These construct definition questions and psychometric issues highlight opportunities for methodological advancement in the measurement of PSC.

Further methodological advancement in the measurement of PSC can be achieved by examining areas of broader organizational safety culture literature that focus on the assessment of culture *strength*. The PSC literature has tended to focus on the level of culture (e.g., is it positive or negative) without directly addressing the issue of culture strength (Marshall et al. 2003). An improved understanding of culture strength is important because strong cultures (e.g., where there is within-group agreement/homogeneity indicating strongly shared perceptions of culture) allow greater behavior predictions by inducing homogeneous expectations regarding accepted behavior while weak cultures offer less reliable predictions of staff safety behavior (Schneider, Salvaggio, and Subirats 2002). Accordingly, there is a need to enhance the understanding of the extent to which staff perceptions of PSC at both the unit and organization levels are truly “shared” (Gaba et al. 2007; Zohar et al. 2007) and there is a need to consider when the aggregation of PSC data may or may not be warranted.

PRESENT STUDY

This paper reports on the continued development of a measure of PSC based on the Modified Stanford Instrument (MSI). The present study's goals included (1) improving the questionnaire's item content and conducting a full psychometric cross-validation of the improved instrument, including investigating its applicability across multiple staff groups and care settings and (2) exploring within-group agreement in PSC measurement. First, we candidly report that this instrument carries similar psychometric challenges to other PSC instruments currently in use. Second, we use this instrument to begin to examine the strength and uniformity of PSC from questionnaire data—an issue that requires attention as we advance the field of PSC measurement (Gaba et al. 2007). We pay particular attention to the degree of within-unit and within-organization agreement (homogeneity) on two key dimensions of PSC (*Organization leadership for safety* and *Unit leadership for safety*) and discuss the implications for measurement and reporting of PSC data in health care settings.

METHODS

The present study was intended to build on psychometric knowledge gained in a previous study of nurse leader perceptions of PSC (Ginsburg et al. 2005). Data from two cross-sectional surveys conducted with a broad range of staff groups in 10 rural and urban Canadian health care organizations were used for this study. The first survey was conducted in fall 2005 with four of the ten organizations as part of a regional effort to assess PSC. The second survey was conducted in fall 2006 with the other six organizations as part of a nationally funded study to examine the psychometric properties of the MSI. The first author led both data collection efforts.

Sample and Questionnaire Administration

All 10 organizations studied provide the full range of clinical services in a variety of care settings (acute, long-term care, community, prehospital care, other). Nine are multisite organizations. In both years, survey methods were similar and the same staff groups and care settings were targeted. *All* direct care providers (nurses, physicians, allied health professionals, and technicians), clinical educators and managers, and support service staff and managers such as unit clerks, housekeeping staff, and health records techs, were targeted to receive a survey. Executive leaders and staff in administrative

departments and research positions were excluded as many survey items are not relevant for these groups. To facilitate data collection, each organization’s HR department provided a list of staff members that included name, job title, unit or department, and site. These lists, together with self-reported job category and care-setting data provided on the questionnaires, were used to assign survey respondents to a staff group, unit/department,² and care setting. This approach allowed us to determine response rates and perform analysis for each of these groups.³

All targeted staff were sent a survey and personalized cover letter, followed by a reminder card 2 weeks later. A second survey was sent to all staff 3 weeks after that.⁴ For the 2005 sample, 5,595 out of 14,108 surveys were returned for a response rate of 40 percent (range of 26–47 percent across organizations). In the 2006 survey, 6,243 surveys were returned out of 22,623 surveys sent out for an overall response rate of 28 percent (range of 18–34 percent across organizations). Table 1 shows response rates for unit type and staff group.

Questionnaire Content

The Modified Stanford PSC Survey Instrument (MSI) is a questionnaire that was adapted from work by Singer et al. (2003) and subsequent work by two

Table 1: Response Rates by Unit* Type and Staff Group

<i>Unit Type</i>	<i>2006 Response Rate (%) (No. of Units)</i>	<i>Staff Group</i>	<i>2005 Response Rate (%) (n)</i>	<i>2006 Response Rate (%) (n)</i>
Patient care units	36.6 (20)	Nurses	40.5 (1,580)	30.4 (2,320)
LTC unit/small site	38.0 (6)	Care aides	28.7 (614)	27.0 (951)
Allied HP department	49.4 (4)	Allied and technicians	48.3 (718)	28.9 (1,164)
Medical/surgical departments	24.4 (4)	Clinical care managers	62.4 (264)	40.6 (365)
Clinical support departments	26.3 (3)	Ward clerks	27.0 (515)	21.6 (270)
All units in Rwg analysis	36.3 (37)	Physicians	49.1 (1,052)	23.6 (386)
		EMS staff	20.9 (108)	21.1 (78)
		Nonclinical managers	56.7 (119)	29.2 (97)
		Nonclinical support staff	45.5 (621)	20.1 (546)
		Other	18.2 (4)	28.7 (66)
		All Staff	39.7 (5,595)	27.6 (6,243)

*Response rates for the 37 units in the 2006 data collection that were used in the within-group homogeneity analysis.

members of the current study team (Ginsburg et al. 2005). In its present form, the MSI also includes four items measuring supervisory leadership for safety currently used in the AHRQ PSC survey (these items were themselves adapted from Zohar (2000)). The fall 2005 (MSI-05) and fall 2006 (MSI-06) had 36 and 38 items, respectively. Minor questionnaire changes made between the MSI-05 and MSI-06 surveys included the removal of a small number of items that were found on the 2005 survey to have low factor loadings and that were not relevant to the dimensions under examination, and the addition of new items designed to capture staff perceptions of patient safety learning behaviors following errors. Both questionnaires also captured data on staff category and care setting and other basic demographics.

Analysis

Following standard psychometric practices for establishing construct validity through a cross-validation study (Van Prooijen and Van Der Kloot 2001), EFA were conducted on the MSI-05 survey data and CFA were conducted on the MSI-06 survey data. When acceptable fit indices (Hu and Bentler 1999) were not achieved, the MSI-06 data were analyzed using a combination of EFA and reliability analyses (Chronbach's α internal consistency, item-analysis [α if item deleted], and test-retest approaches⁵) to attempt to identify a reliable set of dimensions of PSC that would allow us to study the strength of culture issues. To examine the strength of culture, we tested for within-group agreement (homogeneity of responses) by calculating the Rwg coefficients of homogeneity for skewed distributions (James, Demaree, and Wolf 1984) for each survey dimension for the 37 units² with 20+ responders and for the six organizations in the MSI-06 data set. Only MSI-06 data were used for these analyses as the data set was sufficiently large and contained all the same survey items collected using the most recent version of the questionnaire.

RESULTS

The EFA (not shown) conducted on the complete MSI-05 data set suggested a three-factor structure that was consistent with the factor structure we reported previously for nurses in clinical leadership roles (Ginsburg et al. 2005) and that was invariant to staff group and care setting. We were unsuccessful when we tried to confirm this structure in a separate, similar sample⁶ of health care workers by performing CFA on the MSI-06 data set ($n = 4,176$). While all

items loaded significantly on the factors to which they belonged, the model did not yield strong evidence of acceptable fit (Bentler and Bonett 1980). Without using suggested modification indices to retro-fit the model, comparative fit indices, which take sample size into account (Bentler 1990), were below the minimum criteria (CFI = 0.85, GFI = 0.91, NFI = 0.84). Residual-based indices, scaled such that lower is a better fit, showed a mix of good to acceptable fit (SRMR = 0.05, indicating a good fit, and RMSEA = 0.074, indicating borderline fit, [Hu and Bentler 1999]). Splitting the data into staff role and care-setting groups yielded results that ranged from similar to the overall sample (nurses, allied health and techs, adult acute inpatient settings) to results that indicate unacceptable fit on all fit indices (nonclinical support staff, clinical educators, community mental health settings), although the fit was not markedly different for any subgroup.

As noted, the remainder of analysis used only the 2006 data set. We examined the factor structure of the 2006 data using EFA. With the theoretical underpinnings of PSC in mind, these EFA results were examined in tandem with internal consistency reliability data. This process suggested five potential dimensions of PSC. The items in each dimension and their factor loadings from EFA of the MSI-06 data are shown in Table 2. These five factors were derived primarily to allow us to explore the strength of PSC and demonstrate the level of analysis issues. Any reliable factor structure, justifiable on theoretical or empirical grounds, allows us to demonstrate within-and-between unit analysis issues; thus, we chose the set of factors presented in Table 2 over a CFA factor structure retrofitted to match the 2006 sample data with the idea that these factors were theoretically stronger.

The first two dimensions reflect leadership for safety at the organization level and at the unit level. The *Organization leadership for safety* dimension ($\alpha = 0.88$, test-retest $r = 0.82$) has seven items and reflects the extent to which staff perceive that patient safety is valued by senior leadership and is a priority in the organization. The *Unit leadership for safety* dimension ($\alpha = 0.81$, test-retest $r = 0.82$) also has seven items, four of which comprise the AHRQ (Sorra and Nieva 2003) supervisory leadership scale, and three other items that relate to perceptions regarding assessment and management of risks to patients, and rewards for identifying safety problems. Dimensions three and four have lower α s and reflect *Perceived state of safety* (nine items, $\alpha = 0.69$, test-retest $r = 0.74$) and *Shame and repercussions of reporting* related to reporting and talking about errors (four items, $\alpha = 0.69$, test-retest $r = 0.64$). The fifth dimension has five items and reflects *Safety learning behaviors* following major safety events ($\alpha = 0.77$, test-retest $r = 0.76$).

Table 2: Patient Safety Culture Dimensions—Items and Factor Loadings

	<i>Factor</i>				
	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>
Q1. Patient safety decisions are made at the proper level by the most qualified people	0.69	-0.03	0.01	0.01	0.05
Q2. Good communication flow exists up the chain of command regarding patient safety issues	0.68	0.02	0.01	0.03	0.10
Q4. Senior management has a clear picture of the risk associated with patient care	0.66	0.08	0.02	-0.09	0.08
Q7. Senior management provides a climate that promotes patient safety	0.74	0.11	-0.04	-0.04	0.04
Q12. Senior management considers patient safety when program changes are discussed	0.54	0.15	-0.03	-0.04	0.12
Q29. My organization effectively balances the need for patient safety and the need for productivity	0.52	0.15	0.17	-0.10	0.08
Q30. I work in an environment where patient safety is a high priority	0.56	0.11	0.02	0.11	0.05
Q35. Whenever pressure builds up, my supervisor/manager wants us to work faster, even if it means taking shortcuts	0.03	0.45	0.14	0.28	0.01
Q36. My supervisor/manager overlooks patient safety problems that happen over and over	0.14	0.38	0.03	0.31	0.00
Q33. My supervisor/manager says a good word when he/she sees a job done according to established patient safety procedures	-0.06	0.73	-0.04	0.00	0.16
Q34. My supervisor/manager seriously considers staff suggestions for improving patient safety	0.13	0.70	-0.11	0.14	0.05
Q5. My unit takes the time to identify and assess risks to patients	0.57	-0.09	-0.10	0.20	0.12
Q6. My unit does a good job managing risks to ensure patient safety	0.65	-0.10	-0.07	0.18	0.08
Q18. I am rewarded for taking quick action to identify a serious mistake	0.11	0.37	-0.05	-0.04	0.15
Q21. Loss of experienced personnel has negatively affected my ability to provide high quality patient care	0.17	0.10	0.38	0.05	-0.01
Q22. I have enough time to complete patient care tasks safely	0.42	0.16	0.27	-0.03	-0.10
Q24. In the last year, I have witnessed a co-worker do something that appeared to me to be unsafe for the patient in order to save time	0.18	0.03	0.32	0.13	-0.01
Q25. I am provided with adequate resources (personnel, budget, and equipment) to provide safe patient care	0.49	0.16	0.27	-0.13	-0.01
Q26. I have made significant errors in my work that I attribute to my own fatigue	-0.02	-0.02	0.36	0.27	-0.02

continued

Table 2. *Continued*

		<i>Factor</i>				
		1	2	3	4	5
Q27.	I believe that health care error constitutes a real and significant risk to the patients that we treat	0.01	-0.04	0.36	-0.04	0.04
Q28.	I believe health care errors often go unreported	0.05	-0.07	0.42	0.03	0.18
Q11.	I am less effective at work when I am fatigued	-0.01	0.03	0.49	-0.08	0.00
Q13.	Personal problems can adversely affect my performance	-0.09	0.00	0.46	0.07	0.00
Q16.	I will suffer negative consequences if I report a patient safety problem	0.10	0.08	0.02	0.60	0.11
Q3.	Reporting a patient safety problem will result in negative repercussions for the person reporting it	0.16	0.08	0.00	0.55	0.01
Q9.	If I make a mistake that has significant consequences and nobody notices, I do not tell anyone about it	-0.09	-0.01	0.04	0.45	0.03
Q8.	Asking for help is a sign of incompetence	0.06	0.08	-0.02	0.54	-0.03
Q34.	Individuals involved in major events have a quick and easy way to capture/report what happened	0.12	-0.01	0.01	0.11	0.44
Q35.	Individuals involved in major events contribute to the understanding and analysis of the event and the generation of possible solutions	0.13	0.03	-0.04	0.08	0.55
Q36.	A formal process for disclosure of major events to patients/families is followed and this process includes support mechanisms for patients, family, and care/service providers	0.00	-0.01	0.03	-0.01	0.80
Q38.	The patient and family are invited to be <i>directly</i> involved in the entire process of understanding: what happened following a major event and generating solutions for reducing the re-occurrence of similar events	-0.03	0.06	0.08	-0.06	0.64
Q39.	Things that are learned from major events are communicated to staff on our unit using <i>more than</i> one method (e.g. communication book, in-services, unit rounds, e-mails) and/or at <i>several</i> times so all staff hear about it	0.04	0.20	0.04	-0.01	0.50

Extraction method: Principal axis factoring.

Rotation method: Oblimin with Kaiser normalization.

The Rwg analysis of within-group agreement shows stronger within-unit agreement than within-organization agreement on all five dimensions of PSC (see Table 3). Strongest within-group agreement (at both the unit and organization levels) was seen for the *perceived state of safety* dimension, followed by

Table 3: Within-Unit and Within-Organization Homogeneity of Variance

	Median Within-Unit Rwg* n = 37 Units	Median Within-Organization Rwg* n = 6 Organizations
Organization leadership for safety	0.63	0.55
Unit leadership for safety	0.75	0.64
Perceived state of safety	0.78	0.75
Shame and repercussions of reporting	0.75	0.71
Learning behaviors	0.71	0.63

*Based on slight negatively skewed distribution.

the *Unit leadership for safety* and *Shame and repercussions of reporting* dimensions. When examining within-group agreement, it is also typical to assess between-group differences. One-way analysis of variance (*ANOVA*), using unaggregated data in which the between-subject factor was either unit or organization membership (not shown), indicated that all five dimensions showed significant between-unit and between-organization variation, indicating that some units and some organizations are perceived as having a more positive PSC than others on all these dimensions.

DISCUSSION

In this research, we took steps to improve and validate the MSI PSC instrument using widely accepted rigorous psychometric validation procedures. The factor structure that we derived from MSI-05 data did not yield consistently “good” levels of fit according to accepted criteria (Hu and Bentler 1999) when applied to MSI-06 data collected independently in six other organizations. These results are consistent with the fit results for other PSC measures that have either fallen short of achieving required fit levels (Naveh et al. 2005) or have used CFA modification indices to achieve good fit with sample data (Sorra and Nieva 2003; Naveh et al. 2005; Hutchinson et al. 2006; Sexton et al. 2006). While all these efforts are a positive reflection of the numerous initiatives underway to help mature the psychometrics of PSC measurement, the state of PSC measurement is such that none of the currently available tools, including the MSI, have adequate psychometric properties. However, the reasons for this general inability to demonstrate sound PSC psychometrics require further exploration; it could reflect (a) some potential or inherent imprecision in the construct definition of PSC (e.g., the definitional precision of a cloud problem), (b) that the construct of safety culture is highly context

specific as suggested in the organizational literature (Coyle et al. 1995) (e.g., PSC factor structure may be unique to staff group [and differ for nurses, physicians]) and care setting (and differ for ICUs and LTC as an example) or, (c) the need for a more theory-driven construct definition of PSC in health care settings. As previously noted, safety culture has been conceptualized and measured in numerous ways in health care and the broader organizational safety literature, and measures of the *current* construct definition may not be amenable to CFA.

Some of our results lend empirical support to the idea that the current dimensions of PSC may be imprecise as constructs (scenario (a) above). For example, the items “I am provided with adequate resources (personnel, budget, and equipment) to provide safe patient care” and “I have enough time to complete patient care tasks safely” cross-load on the *perceived state of safety* dimension and the *organizational leadership* dimension. These two items, which both have to do with perceptions of how well patient safety is resourced, might reflect organizational leadership for safety *as well as* perceived state of safety in an organization. Two other items that we have included in the *unit leadership* dimension (“My unit takes the time to identify and assess risks to patients” and “My unit does a good job managing risks to ensure patient safety”) actually load on the *organizational leadership* dimension. In addition to suggesting construct imprecision, these two unit-level items with high loadings on the *organizational leadership* dimension may reflect the kinds of mediating and moderating relationships that have been shown to exist between organizational leadership and unit-level leadership for safety (Zohar 2002; Zohar and Luria 2005)—two of the most salient dimensions of safety culture identified in the safety culture literature (Zohar 2000; Flin et al. 2006; Zohar et al. 2007).⁷ Given that the *practice* of PSC measurement may be outpacing the *research*, it is incumbent upon health services researchers to continue to carefully study the measurement properties of this construct and, in the interim, help specify how these measures can and cannot be appropriately used in organizational settings. Implications for PSC assessment in organization are discussed below.

The results of our within group homogeneity analysis indicate that perceptions of PSC on the dimensions we examined are stronger (i.e., are more widely shared or consistent) within *units* than within *organizations*. Given how much more diverse organizations are than units on a number of factors, this result is not surprising. More careful consideration of these results specifically with respect to the *unit leadership for safety* and *organizational leadership for safety* dimensions allows us to highlight some important questions about the most appropriate unit of analysis for these two dimensions. The *unit*

leadership dimension is clearly a group-level construct in that the focus of the dimension is the immediate work area and supervisor. Accordingly, our results showing higher levels of *within-unit* agreement than *within-organization* agreement on this dimension make sense given that organizations are made up of a number of heterogeneous units and supervisors who implement organizational safety policies to different degrees (Westrum 2004). In addition, given that culture data should not be aggregated beyond the level to which the items correspond (Cooper 2000), a PSC dimension that measures *unit leadership* for safety should probably *not* be aggregated to and reported at the organization level.

Decisions regarding appropriate units of analysis for the *organizational leadership for safety* dimension are less clear cut, both theoretically and empirically. Zohar (2000) argues that group-level climate perceptions need to be based on criteria that are relevant to groups, which raises the theoretical question of whether the *organizational leadership* for safety dimension is relevant at the unit level. If we accept that, due to their shared work experiences on a unit, staff members within that unit may have some degree of shared perceptions of senior leadership's prioritization of safety, then it becomes reasonable to aggregate and report data about *organizational leadership for safety* at the unit level. In this instance, the degree of statistical agreement within units provides important information about how widely shared perceptions of *organizational leadership for safety* are at the unit level.

Empirically, if we interpret our data using a consensus approach (Chan 1998) for deciding when to aggregate data for multilevel constructs like safety culture (which argues that only Rwg coefficients > 0.70 are considered to be sufficiently homogeneous for aggregation [Glick 1985]), the units *and* the organizations we studied would lack sufficient homogeneity in their perceptions of *organizational leadership for safety* to warrant aggregation to either of these levels because within-unit and within-organization Rwg values for this dimension were both below the 0.70 cut-off. In contrast, dispersion theorists (Schneider et al. 2002) argue that a consensus approach hides the status of climate strength as a scientific construct and that the lack of homogeneity on this construct is important in and of itself (and well beyond any aggregation decisions) as it indicates the absence of strong, widely shared perceptions regarding *organizational leadership* for safety in the organizations we studied. We suggest it is conceptually meaningful to report unit perceptions of *organizational leadership for safety* and, in keeping with dispersion approaches, it is useful to use information on the degree of within-group agreement to comment on the *strength* of PSC in an organization. With this in mind, we raise one potential methodological issue.

It is important to consider whether low median Rwg values on the *organizational leadership for safety* dimension reflect a truly weak PSC as the dispersion approach suggests. Alternatively, it is possible that low levels of agreement on this dimension reflect a methodological flaw related to poor item design for those questions that reference the “organization” or “senior management.” Poor item design can result from using questions that are open to multiple interpretations or from asking people to respond to questions about which they have insufficient knowledge—two problems that may be more likely in large or highly decentralized organizations like the kind of multisite organizations we studied. For instance, in diffuse organizational structures some staff may interpret questions about “senior leadership” as the organization’s leadership, while others interpret it as their facility’s site leader or even a program leader. Decentralized structures and great physical distances between staff and senior leadership may also leave staff with insufficient knowledge to respond to questions about senior management. We tried to examine these issues and found there was greater agreement on the *organizational leadership for safety* dimension in the only single-site organization in our study (results not shown). Our results encourage us to consider whether low levels of PSC agreement reflect weak cultures or potential interpretation problems with a particular questionnaire. This latter question requires further scrutiny when it comes to measuring perceptions of *organization-level leadership for safety* in large and/or decentralized organizations and additional research is required into who staff see as “senior leadership” and how they define “the organization.” Answers to these important interpretation questions can shed light on the feasibility of measuring *organizational leadership for safety* in large settings and can provide important practical knowledge about simple matters such as what instructions should accompany questionnaires to aid respondents in interpretation.

Health services researchers and decision makers working with PSC data can address some recent criticisms (Marshall et al. 2003) by becoming more cognizant of and reporting on both the *level* of PSC (where high or low refers to the extent to which patient safety is prioritized in an organization or patient care unit) as well as PSC *strength* (where weak or strong reflects the degree of agreement among survey respondents regarding these priorities) (Zohar et al. 2007). Consideration of both level and strength of PSC and consensus and dispersion perspectives can help us to make the most appropriate decisions regarding aggregation and reporting of PSC data. However, it should be noted that in conducting these kinds of unit level analysis, researchers face notable challenges related to how best to attach respondents to units and how to obtain sufficient numbers of responses at the unit level.

This study has some limitations that should be noted. First, the two cross-sectional surveys upon which these analyses are based suffered from low response rates and very little is known about potential nonresponse bias in these kinds of surveys. Low response rates, coupled with the small number of units for which we had sufficient data, allow us to raise questions, but prevent us from drawing deeper conclusions about whether and when it is appropriate to aggregate PSC data. Our low response rates are less problematic for the type of factor and reliability analysis presented here (Hutchinson et al. 2006). The practical implications of low response rates are discussed below. A second limitation pertains to the borderline α s (Nunnally 1978) for two of the dimensions reported here—the *Perceived state of safety* dimension and the *Shame and repercussions of reporting* dimensions both have α s of 0.69. For this reason, we focus our within-group agreement discussion on the two most salient dimensions of PSC (*organizational leadership for safety* and *unit leadership for safety*) that are commonly measured in health care (Flin et al. 2006; Zohar et al. 2007) and other industries (Flin et al. 2000; Zohar 2000) and where we were able to demonstrate strong reliability. Nonetheless, the dimensions of PSC require further study.

In terms of implications for PSC measurement *practice*, it is useful to consider that the science of measurement needs to be far more advanced when used for accountability or external comparison than when used to guide improvement (Solberg, Mosser, and McDonald 1997). Accordingly, we suggest that, while benchmarking is often hailed as one of the leading ways to use PSC data for improvement, the science of PSC measurement is not well developed enough for this type of group comparison and the very public or political implications benchmarking data can have. Low response rates in this and other large-scale survey initiatives further reduce the feasibility of meaningful organizational safety culture comparisons (with low response the data reflect the opinions of a small proportion of staff rather than any meaningful perceptions of PSC). Until researchers are able to specify a more psychometrically sound PSC instrument and unless or until stronger response rates can be achieved, it may be more appropriate for health care organizations wishing to measure PSC (and perhaps meet accreditation and other standards set out in the United States, the United Kingdom, and Canada by JCAHO, the NHS and the CCHSA, respectively) to do any or all of the following.

- (1) Carry out PSC measurement on a smaller set of targeted units or patient care areas where data collection can be undertaken more diligently and response rates closer to 70 percent can and

have been achieved (Kho et al. 2005; Sexton et al. 2006; Zohar et al. 2007).

- (2) Focus on *internal* uses of the data, either (a) comparing relative performance on individual *questions within* organizations to identify opportunities for improvement, and/or (b) comparing performance within organizations over time as response bias is likely to remain fairly constant within organizations.
- (3) Engage in qualitative discussions of survey results to get a sense of how representative an organization's data are and to begin frank discussion of safety in the organization before putting any improvement or change initiatives in place (see Sexton et al. 2007 for a recent example).

Together, these approaches will be reasonably tolerant of any imprecision in the data as well as any systematic or unsystematic bias resulting from low response rates. Accordingly, these approaches will help ensure PSC measurement data that are presently being collected are used appropriately and are meaningful to organizations.

In the meantime, researchers can continue to revise and improve the psychometrics of existing PSC questionnaires and some may also consider revisiting the PSC construct and measurement process from a theory-driven perspective. Other approaches should also be pursued for establishing construct validity through the use of convergent and divergent validity approaches (Singer et al. 2007). This kind of relational approach to establishing construct validity could also be used to examine the relationship between PSC measures and other patient safety outcomes (Singla et al. 2006). Indeed, this was recently done to develop the Safety Organizing Scale, a nine-item behavioral measure grounded in studies of high-reliability organizations (Vogus and Sutcliffe 2007). Finally, future research on the construct of PSC should try to further establish whether dimensions of PSC are invariant to staff group and care setting or whether the construct definition of PSC is population-specific, as suggested in some areas of the organizational safety literature. An answer to this question is important for understanding whether a single PSC questionnaire is likely to meet the needs of large multisite or integrated health systems wishing to assess PSC across the organization, or whether caregiver or setting specific surveys are required.

The current paper examines the psychometrics of one PSC instrument and explores the *strength* of culture and unit of analysis issues. With health care organizations, regulatory, accreditation, and safety agencies quickly adopting

PSC measurement, caution should be exercised until there has been additional careful examination of these important properties of new and existing PSC measures. Such an approach will help ensure there is continued advancement in the related processes of PSC measurement, analysis, interpretation, and data use in health care settings.

ACKNOWLEDGMENTS

Joint Acknowledgement/Disclosure Statement: This study was funded by a grant from the Canadian Patient Safety Institute (CPSI) (RFA-0506132). The lead author (Ginsburg) is also supported by a Career Scientist Award from the Ontario Ministry of Health and Long Term Care (MOHLTC). This research was also supported by in-kind contributions (in the form of costs associated with questionnaire distribution) from each of the 10 organizations that participated in the study. We would also like to acknowledge and thank our key contact people in each of these organizations for helping to facilitate data collection: Mona Pinder and Margaret Sevcik from the Calgary Health Region; Laurie Thompson, Jan Byrd, and Juliet Cummins from the Manitoba Institute for Patient Safety; Pauline MacDonald from Capital Health; Darlene Boliver from the IWK Health Centre; and Nan Brooks from the University Health Network. We thank Evan Castel for helping to code staff group and care setting data.

Disclosures: None.

NOTES

1. It is generally accepted that culture and climate are closely related concepts and that safety climate consists of the surface manifestations of the safety culture and can be measured using quantitative measures. See Schein (1990) and Guldenmund (2007) for a detailed description of the layers of culture. We will use the term patient safety culture (PSC) except where quoting or citing the work of others who use the term climate. We use the more general term "safety culture" when referring to the broader organizational safety literature.
2. The unit/department in most cases reflects the patient care unit where a staff member primarily works. In some cases, the unit refers to small LTC areas (stand alone and attached to acute care centers) where finer unit distinctions are not made within the setting. For staff groups that tend work across larger parts of an organization and are therefore attached to departments rather than units (such as housekeeping staff, allied health professionals, and physicians), this variable

- reflects a department for a particular hospital site (e.g., housekeeping department, physiotherapy department, clinical nutrition department, orthopedics department, department of surgery).
3. We could not determine response rates by care setting as we only had these data from respondents. In our most recent PSC research, we are asking the HR department to also provide data on care setting.
 4. Privacy requirements prevented organizations from releasing staff names to the research team, so organizations mailed out the surveys and surveys were returned directly to the research team. Because it would have violated our ethics protocol to inform organizations which staff members returned a survey, a second survey was mailed to all staff 3 weeks after the reminder card.
 5. Over 400 people completed and returned both surveys they were mailed (see footnote 4), approximately 6 weeks apart. After comparing demographic data to ensure the surveys with the same ID were completed by the same individual, data from these questionnaires were used to establish test–retest reliability.
 6. Both the 2005 and 2006 samples included urban and rural teaching and nonteaching settings. Staff group proportions were similar in the 2005 and 2006 populations and respondent groups with one exception—physicians made up a greater proportion of the 2005 population (18 percent) and respondent group (15 percent) than in 2006 (6 percent of the population and respondent group). Proportions of respondents from each care setting were also similar in 2005 and 2006.
 7. For the purpose of these analyses, these two items were placed with the *unit leadership* items to enhance reliability and face validity of both these dimensions as it is both theoretically and intuitively clearer to have items about unit management of patient safety risks in the *unit leadership* dimension.

REFERENCES

- Bentler, P. M. 1990. "Comparative Fit Indexes in Structural Models." *Psychological Bulletin* 107 (2): 238–46.
- Bentler, P. M., and D. G. Bonett. 1980. "Significance Tests and Goodness of Fit in the Analysis of Covariance Structures." *Psychological Bulletin* 88: 588–606.
- Chan, D. 1998. "Functional Relations among Constructs in the Same Content Domain at Different Levels of Analysis: A Typology of Composition Models." *Journal of Applied Psychology* 83 (2): 234–46.
- Colla, J. B., A. C. Bracken, L. M. Kinney, and W. B. Weeks. 2005. "Measuring Patient Safety Climate: A Review of Surveys." *Quality and Safety in Health Care* 14 (5): 364–6.
- Cooper, M. D. 2000. "Towards a Model of Safety Culture." *Safety Science* 36 (2): 111–36.
- Cooper, M. D., and R. A. Phillips. 2004. "Exploratory Analysis of the Safety Climate and Safety Behavior Relationship." *Journal of Safety Research* 35 (5): 497–512.
- Coyle, I. R., S. D. Sleeman, and N. Adams. 1995. "Safety Climate." *Journal of Safety Research* 26 (4): 247–54.

- Fleming, M. 2005. "Patient Safety Culture Measurement and Improvement: A 'How to' Guide." *Healthcare Quarterly* 8 (special issue): 14–9.
- Flin, R. 2007. "Measuring Safety Culture in Healthcare: A Case for Accurate Diagnosis." *Safety Science* 45 (6): 653–67.
- Flin, R., C. Burns, K. Mearns, S. Yule, and E. M. Robertson. 2006. "Measuring Safety Climate in Health Care." *Quality and Safety in Health Care* 15 (2): 109–15.
- Flin, R., K. Mearns, P. Oapos-Connor, and R. Bryden. 2000. "Measuring Safety Climate: Identifying the Common Features." *Safety Science* 34 (1–3): 177–92.
- Gaba, D. M., S. J. Singer, and A. K. Rosen. 2007. "Safety Culture: Is the Unit the Right Unit of Analysis?" *Critical Care Medicine* 35 (1): 314–6.
- Ginsburg, L., P. G. Norton, A. Casebeer, and S. Lewis. 2005. "An Educational Intervention to Enhance Nurse Leaders' Perceptions of Patient Safety Culture." *Health Services Research* 40 (4): 997–1020.
- Glick, W. H. 1985. "Conceptualizing and Measuring Organizational and Psychological Climate: Pitfalls in Multilevel Research." *Academy of Management Review* 10 (3): 601–16.
- Guldenmund, F. W. 2000. "The Nature of Safety Culture: A Review of Theory and Research." *Safety Science* 34 (1–3): 215–57.
- . 2007. "The Use of Questionnaires in Safety Culture Research—An Evaluation." *Safety Science* 45 (6): 723–43.
- Hofmann, D. A., and B. Mark. 2006. "An Investigation of the Relationship between Safety Climate and Medication Errors as Well as Other Nurse and Patient Outcomes." *Personnel Psychology* 59 (4): 847.
- Hu, L., and P. M. Bentler. 1999. "Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria versus New Alternatives." *Structural Equation Modeling* 6: 1–55.
- Hutchinson, A., K. L. Cooper, J. E. Dean, A. McIntosh, M. Patterson, C. B. Stride, B. E. Laurence, and C. M. Smith. 2006. "Use of a Safety Climate Questionnaire in UK Health Care: Factor Structure, Reliability and Usability." *Quality and Safety in Health Care* 15 (5): 347–53.
- James, L., R. Demaree, and G. Wolf. 1984. "Estimating Within-Group Interrater Reliability with and without Response Bias." *Journal of Applied Psychology* 69: 85–98.
- Katz-Navon, T., E. Naveh, and Z. Stern. 2005. "Safety Climate in Health Care Organizations: A Multidimensional Approach." *Academy of Management Journal* 48 (6): 1075–89.
- . 2006. "The Moderate Success of Quality of Care Improvement Efforts: Three Observations on the Situation." *International Journal for Quality in Health Care* 19 (1): 4–7.
- Kho, M. E., J. M. Carbone, J. Lucas, and D. J. Cook. 2005. "Safety Climate Survey: Reliability of Results from a Multicenter ICU Survey." *Quality and Safety in Health Care* 14 (4): 273–8.
- Marshall, M., D. Parker, A. Esmail, S. Kirk, and T. Claridge. 2003. "Culture of Safety." *Quality and Safety in Health Care* 12 (4): 318.

- Naveh, E., T. Katz-Navon, and Z. Stern. 2005. "Treatment Errors in Healthcare: A Safety Climate Approach." *Management Science* 51 (6): 948.
- Nieva, V. F., and J. Sorra. 2003. "Safety Culture Assessment: A Tool for Improving Patient Safety in Healthcare Organizations." *Quality and Safety in Health Care* 12 (suppl 2): ii17–23.
- Nummally, J. C. 1978. *Psychometric Theory*, 2d Edition. New York: McGraw-Hill.
- Pace, W. D. 2007. "Measuring a Safety Culture: Critical Pathway or Academic Activity?" *Journal of General Internal Medicine* 22 (1): 155–6.
- Pronovost, P., C. G. Holzmueller, D. M. Needham, J. B. Sexton, M. Miller, S. Berenholtz, A. W. Wu, T. M. Perl, R. Davis, D. Baker, L. Winner, and L. Morlock. 2006. "How Will We Know Patients Are Safer? An Organization-Wide Approach to Measuring and Improving Safety." *Critical Care Medicine* 34 (7): 1988–95.
- Pronovost, P., and B. Sexton. 2005. "Assessing Safety Culture: Guidelines and Recommendations." *Quality and Safety in Health Care* 14 (4): 231–3.
- Reason, J. 1997. *Managing the Risks of Organizational Accidents*. Aldershot, UK: Ashgate.
- Schein, E. H. 1990. "Organizational Culture." *American Psychologist* 45 (2): 109–19.
- Schneider, B., A. N. Salvaggio, and M. Subirats. 2002. "Climate Strength: A New Direction for Climate Research." *Journal of Applied Psychology* 87 (2): 220–9.
- Sexton, J. B., R. L. Helmreich, T. B. Neilands, K. Rowan, K. Vella, J. Boyden, P. R. Roberts, and E. J. Thomas. 2006. "The Safety Attitudes Questionnaire: Psychometric Properties, Benchmarking Data, and Emerging Research." *BMC Health Services Research* 6: 44.
- Sexton, J. B., L. A. Paine, J. Manfuso, C. G. Holzmueller, E. A. Martinez, D. Moore, D. G. Hunt, and P. J. Pronovost. 2007. "A Check-Up for Safety Culture in My Patient Care Area." *Joint Commission Journal on Quality and Patient Safety/Joint Commission Resources* 33 (11): 699, 703, 645.
- Singer, S. J., D. M. Gaba, J. J. Geppert, A. D. Sinaiko, S. K. Howard, and K. C. Park. 2003. "The Culture of Safety: Results of an Organization-Wide Survey in 15 California Hospitals." *Quality and Safety in Health Care* 12 (2): 112–8.
- Singer, S. J., M. Meterko, L. Baker, D. M. Gaba, A. Falwell, and A. K. Rosen. 2007. "Workforce Perceptions of Hospital Safety Culture: Development and Validation of the Patient Safety Climate in Healthcare Organizations Survey." *Health Services Research* 42 (5): 1999–2021.
- Singla, A. K., B. T. Kitch, J. S. Weissman, and E. G. Campbell. 2006. "Assessing Patient Safety Culture: A Review and Synthesis of the Measurement Tools." *Journal of Patient Safety* 2 (3): 105–15.
- Solberg, L. I., G. Mosser, and S. McDonald. 1997. "The Three Faces of Performance Measurement: Improvement, Accountability, and Research." *Joint Commission Journal on Quality Improvement* 23 (3): 135–47.
- Sorra, J., and V. F. Nieva. 2003. *Psychometric Analysis of the Hospital Survey on Patient Safety*. Technical Report Delivered to the Agency for Healthcare Research and Quality (AHRQ). AHRQ Publication No. 04-0041. Rockville, MD: Agency of Healthcare Research and Quality.

- Van Prooijen, J., and W. A. Van Der Kloot. 2001. "Confirmatory Analysis of Exploratively Obtained Factor Structures." *Educational and Psychological Measurement* 61 (5): 777.
- Vogus, T. J., and K. M. Sutcliffe. 2007. "The Safety Organizing Scale: Development and Validation of a Behavioral Measure of Safety Culture in Hospital Nursing Units." *Medical care* 45 (1): 46–54.
- Westrum, R. 2004. "A Typology of Organisational Cultures." *Quality and Safety in Health Care* 13 (suppl 2): ii22–7.
- Zohar, D. 2000. "A Group-Level Model of Safety Climate: Testing the Effect of Group Climate on Microaccidents in Manufacturing Jobs." *Journal of Applied Psychology* 85 (4): 587–96.
- . 2002. "Modifying Supervisory Practices to Improve Submit Safety: A Leadership-Based Intervention Model." *Journal of Applied Psychology* 87 (1): 156.
- Zohar, D., Y. Livne, O. Tenne-Gazit, H. Admi, and Y. Donchin. 2007. "Healthcare Climate: A Framework for Measuring and Improving Patient Safety." *Critical Care Medicine* 35 (5): 1312–7.
- Zohar, D., and G. Luria. 2005. "A Multilevel Model of Safety Climate: Cross-Level Relationships between Organization and Group-Level Climates." *Journal of Applied Psychology* 90 (4): 616–28.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article:

Appendix SA1: Author Matrix.

Appendix SA2: Other Contributions.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.