2013-05-01

# Assessing Adult Medicine Specialists Using Multisource Feedback: A Longitudinal Study

Hurd, Carmen

UNIVERSITY OF CALGARY

# Assessing Adult Medicine Specialists Using Multisource Feedback: A Longitudinal Study

by

Carmen Hurd

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF MEDICAL SCIENCES

CALGARY, ALBERTA

APRIL, 2013

## Abstract

The Physician Achievement Review is a multisource feedback program of practicing physicians that intends to assess a wide range of professional competencies. This longitudinal study focused on the reliability and validity of the PAR assessment in a large sample of adult medicine specialists (n = 404).

Scores on all surveys were high and negatively skewed. All surveys had high internal consistency reliability and moderate generalizability. The three to four factor solutions proposed at Iteration 1 provided for moderate model fit using confirmatory factor analysis at Iteration 2. Scores increased over time, but the effect sizes were small to moderate. There was little or no correlation between self-assessment and medical colleagues on corresponding attributes.

Future research should focus on decreasing score inflation, improving the internal structure of the surveys, and understanding factors that influence score improvements over time.

## Table of Contents

## List of Tables

## Chapter 1: Introduction

There is a growing demand for physician accountability, patient safety, and continuous quality improvement in health care (Norcini, 2005). Not only are doctors expected to have sound medical knowledge and clinical competency, but they are also expected be proficient in non-cognitive domains. In Canada, these expectations are captured by the CanMEDS competency-based framework, which identifies seven physician roles that lead to optimal health and health care outcomes: medical expert (the central role), communicator, collaborator, manager, health advocate, scholar, and professional. These roles are in keeping with the six core competencies of the Accreditation Council for Graduate Medical Education in the United States, and the General Medical Council's (GMC) "Good Medical Practice" in the United Kingdom. These physician competencies have been integrated not only into medical school and residency programs, but also into maintenance of certification for practicing physicians. For example, every five years physicians in the UK participate in "Revalidation", where they are required to demonstrate proficiency in all core principles and values outlined by GMC's Good Medical Practice.

Stemming from the prevailing expectation of proficiency in various competencies, the exigency for tools to assess these competencies arises. Increasingly, multisource feedback (MSF) is emerging as a useful tool in evaluating a range of physician attributes. MSF - also referred to as $360^{o}$ evaluation - involves a systematic collection of feedback by those people with whom the individual being assessed interacts on a routine basis (eg., supervisors, coworkers, clients/patients). Internationally, it is emerging as an important assessment process of practising physicians. For example, in Canada it is currently used by three provincial regulating bodies to assess practicing physicians, and is a key component of revalidation in the UK.

The Physician Assessment Review (PAR) is a MSF assessment program developed through joint collaboration between the College of Physicians and Surgeons of Alberta

(CPSA) and the Universities of Calgary and Alberta (Hall et al., 1999). Participation in the program has been mandatory since 1999, and all licensed physicians in the province complete a full assessment every five years. More recently, it has been adopted by the Colleges of Physicians and Surgeons of Nova Scotia and of Manitoba. The assessment is comprehensive and is intended to cover a broad range of attributes, including medical competence, office management, communication, collegiality, and psychosocial management.

Surveys specific to different medical specialties have been developed. For adult medicine specialists, a complete assessment consists of surveys completed by 25 patients (40 items per survey), 8 non-physician coworkers (22 items per survey), 8 physician colleagues (38 items per survey), and a self-assessment (37 items per survey). The self and the medical colleague survey are identical, with the exception of the item "If a member of my own family needed care I would rate this physician", which is not on the self-survey. Each of the survey items are scored on a 1 to 5 point Likert scale (1 = strongly disagree, 5 = strongly agree), with the option of selecting "unable to assess". Data is collected, analyzed, and reported by an independent research firm called "Pivotal Research Inc" (PAR website). The four surveys used for adult medicine specialists are in Appendix A-D.

After completion of the entire assessment, each participant is given a detailed structured profile of his/her results on each item and on each attribute, along with comparisons to other physicians in the same specialty. Flags identify personal items/attribute scores that are < $10^{th}$ percentile or > 90% percentile compared to the reference norm. The profile includes some suggested steps to encourage reflection and self-improvement. The assumption is that physicians will use the feedback profile to make positive performance changes.

Evidence for reliability and validity of the PAR surveys has been evaluated in different medical specialties, including adult medicine (Violato, Lockyer, Toews & Fidler, 2003;

Lockyer & Violato, 2004), pathologists (Lockyer, Violato, Fidler & Alakija, 2009), psychiatrists (Violato, Lockyer, & Fidler, 2008a), radiologists (Lockyer, Violato, & Fidler, 2008), emergency room physicians (Lockyer, Violato, & Fidler, 2006a), pediatricians (Violato, Lockyer, & Fidler, 2006), and anesthesiologists (Lockyer, Violato, & Fidler, 2006b). These studies consistently find high internal reliability (Cronbach's alpha range from 0.93 to 0.99) and dependability of the overall process (generalizability coefficients range from 0.56 to 0.88). Factor analyses of survey items yield 2-5 factor solutions for all subspecialties that consistently explain > 60% of the variance.

The present study aims to build on the existing foundation of validity-related evidence to support the PAR assessment of adult medicine specialists. It will also inform the current applicability of the surveys, which have been unchanged since inception. To date, the validity-related evidence comes from a small original sample size (n = 103) of adult medicine specialists and was completed almost a decade ago. The strength of this study is its large sample size (n = 404), the use of current data, and the longitudinal nature of the study.

**The following research questions will be addressed in the current study:**

1. How do coworkers, patients, and medical colleagues rate adult medicine specialists on various items and attributes?
2. What underlying attributes does each survey actually measure, and are these stable over time?
3. Are the current surveys reliable?
4. Do scores improve over time, and if yes, can predictors of those changes be identified?
5. What is the relationship between self-assessment attribute scores and corresponding medical colleague attribute scores?

This thesis is divided into five main chapters. This chapter introduced the reader to the PAR assessment program and to the purpose of the current study. Chapter 2 – Literature Review – will present relevant background theory and evidence to understand the importance of the current study. Chapter 3 – Methods – will give a detailed description of the statistical analyses used to answer each research question. Basic explanations of these statistical techniques are also provided. Chapter 4 – Results – presents data related to each research question for each survey consecutively. Finally Chapter 5 – Discussion – will compare our results with past research, discuss their practical implications, and suggest areas for future research.

# Chapter 2: Review of the Literature

## 2.1 An Introduction to Validity

The research questions addressed in the current study relate directly to the underlying validity of the PAR assessment process. The most recent Standards for Educational and Psychological Testing (1999) defines validity as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests", with the proposed interpretation referring directly to the proposed constructs (physician attributes in the current study). More simply, validity refers to how well we can trust the interpretation of results for a given assessment (Cook & Beckman, 2006). Validity is a unitary concept, but evidence from five different sources inform overall validity of an assessment. Concerning the following five sources of validity, greater evidence becomes essential as the stakes of the assessment rise.

- *Content.* Refers to the relationship between the items in the instrument and the attributes it intends to measure. This aspect of validity will not specifically be addressed in the current study.
- *Response processes*. Refers to "evidence of data integrity such that all sources of error associated with the test administration are controlled or eliminated to the maximum extent possible" (Downing, 2003). It refers not only to the actual responses of ratees, but also to the appropriateness of the methods used to combine various types of scores, and the usefulness and accuracy of the score reports provided to examinees (Downing, 2003). This aspect of validity will be addressed in the current study through examination of the descriptive statistics of items and scales (Research Question 1).
- *Internal structure*. Refers to the reliability and factor structure of the assessment. This aspect of validity is explored in detail in the current study. Specifically, the factor structure will be explored at iteration 1 (using exploratory factor analysis) and longitudinally (using confirmatory factor analysis) (Research Question 2). These analyses will clarify how well PAR is actually measuring the proposed attributes.

Additionally, evidence of reliability of the surveys and generalizability of the assessment process will be determined (Research Question 3).

- *Relationship to external variables*. This aspect of validity refers to the relationship with other instruments assessing the same attribute. This will not be explored in the current study. However, evidence has emerged in the last two years that some MSF programs are not able to identify poorly performing trainees (Mitchell, Bhat, Herbert, & Baker, 2011) or practising physicians (Archer & McAvoy, 2011).
- *Consequences* – This aspect of validity refers to the overall impact of the assessment process. This aspect of validity will be addressed by determining change over time (Research Question 4).

### 2.2 Multisource Feedback in Physicians: "Landmark" Studies

One of the first studies to explore the use of MSF in the medical professions was conducted by Linn, Oye, Cope, and DiMatteo (1986), who used MSF to evaluate humanistic qualities of internal medicine residents and faculty in an outpatient clinic setting. Evidence of feasibility and reliability was obtained using ten non-physician coworkers (two nurses, three nurses aide's, one social worker, and four clerical workers) and seven consecutive patients. Additionally, trainee assessments by patients, coworkers, and physician supervisors were moderately correlated, suggesting measurement of the same underlying attribute. Later, Butterfield and Mazzaferri (1991) reported that nurses can reliability assess the humanistic skills (respect, integrity, compassion) of internal medicine residents. Furthermore, moderate positive correlations were found between nurse ratings and those of attending faculty and the evaluation committee, providing validity evidence to support the relationship with external variables.

To our knowledge, the earliest publication exploring the use of MSF to assess practising physicians (rather than trainees) was conducted by Ramsey et al. (1993), who used medical colleague peers to evaluate performance in humanistic, communication, and clinical skills of practicing internal medicine specialists. Eleven raters were needed to achieve a generalizability coefficient > 0.7. Using principal component analysis with

varimax rotation, a two factor solution was obtained - cognitive/clinical skills and humanistic/psychosocial skills – which accounted for 89.5% of the variance. It is worth noting that in this study, peer raters were selected either by the physician ratees (self-selected raters) or by the ratee's supervisor (assigned raters). Ratings did not differ between these two rater groups, suggesting that self-selection of assessors did not undermine the validity of the assessment. More recently, however, using the Sheffield Peer Rating Assessment Tool, Archer and McAvoy (2011) found that 50% of assigned peer raters gave scores of "less than satisfactory" to physicians previously identified as performing poorly, but this dropped to 19% when raters were self-selected. Thus, due to its potential for response bias, the appropriateness of rater self-selection in MSF remains controversial.

As discussed in Chapter 1, specialty-specific PAR surveys were developed and refined over the last decade. These studies consistently found high scores on all surveys, high internal reliability (Cronbach's alpha range from 0.93 to 0.99) and dependability of the overall process (generalizability coefficients range from 0.56 to 0.88). Factor analyses of survey items yield 3-5 factor solutions for all surveys and subspecialties. The underlying attributes include clinical competency, as well as other CanMEDS competencies such as communication and professionalism.

The largest study to date using MSF to assess practicing physicians was recently published (Wright et al., 2012), and is presented here to allow direct comparison of reliability and validity-related evidence with PAR research. The study focused on the UK's General Medical Council's (GMC) patient survey and medical colleague surveys, both of which are shorter than the corresponding PAR survey and each give the option of written comments. The GMC patient questionnaire has 11 items (versus 40 for PAR) and the GMC medical colleague questionnaire has 19 items (versus 37 for PAR). Participation in this study was voluntary, and a range of physician specialties/practice settings were represented. The strength of this study was the large sample size: 1065 practicing physicians participated, which represented 30333 patient ratings (patients were

recruited consecutively), and 17012 colleague ratings (self-selected raters, 10 physicians, 10 non-physician coworkers per participant). Consistent with previous PAR studies, item scores for both questionnaires were high and negatively skewed, and questionnaire internal consistency reliability was high (Cronbach's alpha 0.87 for patient survey and 0.94 for medical colleague survey). Using generalizability theory (raters nested within doctors, crossed with items), 34 patients and 15 colleagues were required to achieve G coefficients > 0.7. Using principal component analysis with varimax rotation, and Kaiser's rule for factor retention, two factor solutions were found for both surveys. Patients who identified their visit as "very important" were more likely to give high ratings. Colleagues who had contact with the physician frequently (most days) were more likely to give high ratings.

These "landmark" studies suggest that MSF assessment of physicians is feasible, reliable, and can assess non-cognitive attributes such as communication and professionalism. Across different assessments, scores are consistently high, which may represent a response bias. In contrast to non-PAR assessments which consistently find 2 factor solutions with EFA, PAR assessments yield 3-5 factor solutions across surveys. The reason for this discrepancy remains unclear. Of the above-mentioned studies, Ramsey et al. (1993) was the only study to focus exclusively on adult medicine specialists. As assessment findings may not generalize across specialties, and as the current study focuses on adult medicine specialists to the exclusion of all others, an up-to-date review of PAR assessment data from adult medicine specialists is presented next.

### 2.3 PAR Assessment in Adult Medicine Specialists

Preliminary evidence to support the reliability and validity of the PAR assessment for adult medicine specialists has previously been reported. The one published article focused exclusively on the medical colleague survey among a small sample (n = 103) of adult medicine specialists (Lockyer & Violato, 2004). Given the impractically of developing different specialty-specific surveys, the underlying impetus was to assess whether or not a single generic peer survey could be used across three specialties (adult

medicine specialists, pediatrics and psychiatry). A working group of experts was recruited to develop the questionnaire. Internal consistency was high for the overall survey and for each attribute score (Cronbach's alpha > 0.9 for all). A mean of 7.6 raters gave a generalizability coefficient of 0.82. Four factors accounted for 73.4% of the variance: Patient Management, Clinical Assessment, Professional Development, and Communication.

Further evidence to support the reliability and validity of the PAR assessment for adult medicine specialists was enumerated in a technical report submitted to the Alberta CPSA (Violato et al. 2003). The sample (n = 103) was identical to that of Lockyer & Violato (2004). The following is a summary of the key findings for the coworker, patient, and self surveys. The findings of the medical colleague survey were also reported in the technical report, but are not presented here because they were identical to the findings reported by Lockyer & Violato (2004).

- Mean ratings for all items were high – greater than 4/5 for all four surveys - indicating a potential response bias for patients, coworkers, and medical colleagues. This finding is consistent with previous PAR-related publications and of those of Wright et al. (2012).
- *Coworker Survey*. Cronbach's alpha was 0.96, indicating very high internal consistency. A three factor solution explained 66.6% of the variance: Professional Communication, Co-worker Collegiality, and Humanistic/Psychosocial.
- *Patient Survey*. Cronbach's alpha was 0.99, indicating very high internal consistency. Fourteen items had levels of "unable to assess" greater than 20%. Five factors were identified that explained 79.1% of the variance: Humanistic, Technical Communication, Personal Communication, Staff, and Office.
- *Self Survey*. Cronbach's alpha was 0.98, indicating very high internal consistency. A four factor solution explained 72.7% of the variance: Psychosocial Management of Patients, Clinical Performance, Humanistic and Communication, and Professional Self-Management.

These two studies provide preliminary evidence to support the reliability and factor structures of the PAR assessment for adult medicine specialists. The current study aims to add to this existing framework in several ways. First, the larger sample size in the current study is beneficial. Secondly, the longitudinal nature of the study will allow a more detailed analysis of the factor structure by allowing its stability over time to be assessed. Additionally, the longitudinal nature of the current study allows us to investigate for change over time. If present, improvement in scores will suggest an educational impact of PAR, providing evidence to support its consequential validity.

### 2.4 What Physician Attributes does PAR Actually Measure?

The current study proposes to clarify the physician attributes that each survey actually measures (Research Question 2). As discussed previously, analysis of PAR data consistently yields 3-5 underlying attributes. Interestingly, these findings are not in keeping with non-PAR MSF tools, which consistently yield two factor solutions. For example, Ramsey et al. (1993), Wright et al. (2012), Archer et al. (2008) and Archer, McGraw and Davies (2010) all reported two factor solutions: typically one factor measuring clinical competencies and one measuring non-clinical competencies. Overall, these non-PAR studies reported higher percentage variance accounted for and higher pattern coefficients in the EFAs, compared to the current study.

Factor analysis is a statistical method that can be used to inform validity of the internal structure of the surveys. Specifically, it is used to determine if the items intended to measure a given construct (in this case, a specific physician attribute) are actually measuring that construct. A primary goal is to explain the most item variance in the fewest number of underlying latent constructs. For example, items intended to measure an aspect of attribute A should have high correlation with other items measuring attribute A, and lower correlations with items intended to measure unrelated constructs. A hypothesized model of relationships between items with underlying attributes can then be confirmed on a second data set to determine how well the model fits the underlying data.

The largest published longitudinal study using PAR data examined the stability of the factor structure and change in performance over time in 250 family medicine physicians who participated in PAR on two occasions, five years apart (Violato, Lockyer & Fidler, 2008b). It is the first and only PAR study to date to use confirmatory factor analysis (CFA) to assess the stability of factors over time. CFA allows one to test the hypothesized interrelationships between measured variables (items on the PAR surveys) and latent variables (physician attributes) (Violato & Hecker, 2007). In this study, underlying attributes were proposed based on an exploratory factor analysis at iteration 1. This model can then be used at Iteration 2 to see how well it "fits" (explains) the data. Comparative fit indices were 0.91 for the medical colleague survey, 0.87 for coworkers, and 0.81 for patient data. Although these values do not meet the convention criteria of > 0.95 to accept the model as having a "good fit" (Hu & Bentler, 1999), they are still acceptable given the long time between iterations and the complexity of the proposed models.

Finally, as highlighted by Archer and McAvoy (2011), scores awarded by different rater groups tend not to correlate, despite claiming to measure the same underlying attribute. For example, the assessment of "communication" by patients may not correlate with that of coworker raters. The lack of correlation suggests two different underlying attributes are being measured. This lack of correlation between rater groups for similar attributes is most defined for the patient raters. Evans, Edwards, Evans, Elwyn and Elwyn (2007) systematically identified and reviewed six surveys used for patient assessment of physicians and found little data available on correlation with other attribute assessments. Thus it seems that different rater groups provide different perspectives on similar underlying attributes.

### 2.5 Does MSF Improve Physician Performance?

One of the main purposes of feedback is to promote learning (Norcini & Burch, 2007). An underlying assumption of MSF is that it has educational impact; that is, the feedback will be used to improve performance in one or more practice areas. For example, the

PAR website states that "The unbiased feedback is enormously helpful to doctors, who will be able to build on their strengths and correct any possible problems" and that "through this program, individual physicians will be able to … institute changes to medical practice that will improve health care for all Albertans". Despite this well-accepted assumption, there is no compelling evidence to date that MSF has an educational impact among practicing physicians.

A major barrier in demonstrating improved performance in practising physicians is the limited opportunities for ongoing evaluation and assessment. The majority of studies in this area focuses on residents, presumably because of the ongoing opportunities for assessment in this group. Cope, Linn, Leake and Barrett (1986) studied the effect of feedback of patient ratings on the performance of internal medicine residents. Residents with the lowest scores from patient ratings were randomized to receive feedback and tailored teaching in the form of suggestions for improvements (intervention group) versus no feedback or teaching (control group). Repeat assessment six months later showed significantly more improvement in the intervention group. Unfortunately, it is unknown if this improvement is due to MSF, to tailored teaching, or to the awareness of being in the interventional group. More recently, Brinkman et al. (2007) randomized 36 first year pediatric residents to receive MSF from nurses and patients' parents, combined with a tailored coaching program (intervention group) or to receive standardized feedback only (control group). In repeat assessments five months later, residents in the intervention group showed greater improvements from baseline in nurse-rated communication and professional behaviours compared to the control group. Both of these studies suggest an educational effect of MSF, however, the available data is limited and may be prone to publication bias. Additionally, it unknown if results from trainees can be generalized to more experienced practising physicians.

In a PAR longitudinal study of 250 family medicine physicians, Violato, Lockyer and Fidler (2008) reported a significant increase in overall medical colleague score between iterations, and the effect size was moderate (Cohen's d = 0.66). The sole unique

predictor of medical colleague score at Iteration 2 was years in practice, although it accounted for only 2.1% of variance. Overall coworker scores increased significantly over time, but the effect size was small (Cohen's d = 0.22). There was no significant increase in patient scores over time. The authors' proposed explanations to explain the small/lack of improvement in scores over time include ceiling effect of the scores, or that the process is insensitive to detect large changes in one or two areas/items.

Due to the scarcity of *objective* evidence, we will review the available *subjective* evidence of the educational impact of MSF. In a survey three months after receiving PAR feedback , 83% of responding family physicians contemplated a change in at least one practice domain, and 66% already initiated a change. Change was more likely to occur in response to patient feedback, particularly if it was a domain over which the physician had control (such as communication) (Fidler, Lockyer, Toews & Violato, 1999). Initiation of change tended to be most common in physicians with lower scores, perhaps suggesting an educational effect. In another study, Sargeant et al., (2003), found that 61% of family physicians indicated they had, or would, initiate practice change in response to participation in the PAR program, particularly in the area of communication. Of the three rater groups, physicians thought the patients were most accurate. They were more likely to agree with the medical colleague ratings if they were high than if they were low. Similarly, internal medicine specialists participating in a voluntary MSF program as part of the American Board of Internal Medicine's Continuing Professional Development Program (ABIM program) felt that the addition of peer and patient feedback had educational value (Lipner, Blank, Leas & Fortna, 2002): 82% indicated they would continue to seek feedback from patients and peers, and 65% percent of participants indicated it would help them to improve the quality of care they provide**.**

In contrast, an equal number of studies have found that doctors perceive MSF to have low educational value. Lockyer, Violato and Fidler (2003) surveyed surgeons three months after participating in PAR to determine the likelihood of implementing change based on their assessment report. Overall, surgeons indicated low likelihood of implementing

change on a broad range of medical competencies; mean of all items was less than 2.30 on a 5 point Likert scale (1 = not considering implementing, 5 = very likely to implement). Furthermore, Murphy, Bruce, Mercer and Eva (2009) found that MSF ranked the lowest of six different work-place based assessments of postgraduate trainees in the UK in terms of participants' perceived educational impact.

Program-specific characteristics likely influence the educational impact of MSF in practising physicians. This was best demonstrated by Overeem et al. (2010), who directly compared three established MSF programs among consultants from varied specialties in hospital-based settings in the Netherlands. The programs included the PAR program (n = 45), the ABIM program (n = 30), and the Dutch Appraisal and Assessment Survey (AAI) (n = 45). The latter is purely qualitative in nature, where colleagues and coworkers are asked to list three strengths and give three suggestions for improvement. These comments are then summarized and fed back to the consultant. Participants in all three programs were interviewed with a trained facilitator to review the MSF report. Of the three MSF programs, consultants viewed PAR as the least satisfying (AAI 89% > ABIM 75% > PAR 53%). However, the majority of consultants expressed intention to change in response to AAI feedback (66%) and PAR feedback (61%), compared to a minority in response to ABIM feedback (25%).

The perceived educational impact of MSF manifests several conflicting findings. Several potential explanations for this exist. First, as demonstrated by Overeem et al. (2010), program-specific characteristics likely influence physicians' acceptance of the feedback. Second, it is possible that acceptance of feedback and willingness to change varies between residents and practising physicians, and between different medical specialties (for example, family physicians may inherently be more open to feedback than surgeons). Finally, evidence suggests that feedback is more accepted when participation in the assessment process is voluntary, rather than mandatory. For example, Lockyer et al. (2011) found that physicians are more open to feedback that originated from activities of

their choosing (such as non-formal meetings/discussions with peers), and more resistant to feedback from mandatory participation (such as the PAR program).

### 2.6 Physician Self-Assessment

In the medical professions, self-assessment has been defined as "a personal evaluation of one's professional attributes and abilities against perceived norms" (Colthart et al., 2008). The self-assessment is one of the four surveys included in a PAR assessment for adult medicine specialists. Although the actual purpose of the self-assessment is not stated explicitly on the PAR website or on the feedback report, one would assume that score gaps between self-assessment scores and medical colleague scores on the same items or attributes would heighten a participant's desire for change. However, a recent comprehensive review emphasised that the educational impact of self-assessment is unknown. In fact, no published studies have explored the effect of self-assessment on actual changes in clinic practice or patient outcomes (Colthart et al., 2008). Paradoxically, the ability to self-assess is an underlying assumption of most continuing professional development programs (Lockyer, Violato & Fidler, 2007), and is considered essential to professional self-regulation (Eva & Regehr, 2005).

Evidence suggests that physicians are inaccurate at self-assessment. For example, in a recent systematic review, two-thirds of studies found little, no, or an inverse relationship between physician self-assessment and external observations of performance (Davis et al., 2006). Those who perform the poorest tend to be the least accurate at self-assessment (Colthart et al., 2008). Violato and Lockyer (2006) found a similar trend with PAR: physicians in the top quartile of peer ratings tended to self-rate themselves 30-40 percentile ranks lower than their medical colleague peers, whereas physicians in the lowest quartile of peer ratings tended to rate themselves 30-40 percentile ranks higher than their peers. This discrepancy between self and peer assessment was consistent across three disciplines (psychiatry, pediatricians and adult medicine specialties), and across attribute scores. Thus a potential dilemma emerges in that those physicians with the greatest need for improvement may be the least receptive to negative feedback, as it

may differ from their self-perception. If true, this may decrease the effectiveness of MSF. Those who receive negative feedback may not change because they view the assessment as inaccurate, and those who receive positive feedback may not change because the assessment confirms that they are doing well.

In a longitudinal study using PAR family physician data, Lockyer et al. (2007) attempted to identify predictors of self-assessment scores at Iteration 2. Using a paired t-test, the mean total score at Iteration 2 was significantly higher than at Iteration 1, but the effect size was only moderate (Cohen's d = 0.46). Two variables, Professionalism/Communication Score and Psychosocial Score from the Self survey at Iteration 1, explained 27.4% of the variance of the total score at Iteration 2. However, demographic factors or attribute scores from the other three surveys did not predict scores at Iteration 2. These findings suggest that self-assessment scores of family physicians tend to be stable over time, and do not seem to be influenced by feedback from physician peers, coworkers, or patients. It is not known if the finding of stability of self-assessment scores in family physicians can be generalized to other specialties. The current study will determine if self-assessment scores change between iterations, and if predictors of change can be identified.

### 2.7 Summary and Research Questions

In summary, the current study hopes to build on the existing foundation of validity-related evidence in support of the PAR assessment of adult medicine specialists. To date, evidence to support the reliability and evidence of validity comes from a small sample size (n = 103) of adult medicine specialists, and was completely approximately a decade ago. The strength of the current study is its large sample size (n = 404), the up-to-date data, and the longitudinal nature of the study.

The following questions will be addressed in the current study:

1. How do coworkers, patients, and medical colleagues rate adult medicine specialists on various items and attributes? These ratings may provide insight into potential

response biases in various rater groups. Based on previous PAR research, it is hypothesized that ratings will be high and negatively skewed.

2. What underlying attributes does each survey actually measure, and are they stable over time? The current study will explore evidence to support internal structure validity using both EFA and CFA. A proposed model (relationships between items and attributes) derived using EFA at Iteration 1 will be tested using data at Iteration 2. If our proposed model accounts for rater responses, and if this structure is stable over 5 years, it will lead to good model fit at Iteration 2.
3. Are the current surveys reliable? This question has practical implications for implementation of PAR.
4. Do scores improve over time, and if yes, can predictors of those changes be identified? The current study will also attempt to assess the educational impact by exploring for change in scores between iterations. Based on the underlying assumption that PAR has an educational impact, we expect scores to increase over time. We also intend to identify what factors predict this increase.
5. What is the relationship between self-assessment attribute scores and corresponding medical colleague attribute scores? Currently, little is known about this relationship in adult medicine specialists.

# Chapter 3: Methods

A longitudinal study was conducted using Alberta PAR data from adult medicine specialists who participated on two occasions, 5 years apart, between 1999 and 2010. This chapter will begin with a description of the PAR surveys. Next, participant selection and data collection is described. Finally, the methods of data analysis for each research question will be specified in detail.

## 3.1 Description of the Surveys

A complete assessment consisted of surveys by 25 patients (40 items per survey), 8 non-physician coworkers (22 items per survey), 8 physician medical colleagues (38 items per survey), and a self-assessment (37 items per survey). These surveys are shown in Appendix A to D. The self and the medical colleague survey are identical with the exception of one item "If a member of my own family needed care I would rate this physician", which is not on the self-survey. Each of the survey items are scored on a 1 to 5 point Likert scale (1 = strongly disagree, 5 = strongly agree), with the option of selecting "unable to assess". Data are collected, analyzed, and reported by Pivotal Research Inc, a private research company hired by the College of Physicians and Surgeons of Alberta. Previous work gives preliminary evidence to support the reliability and validity of these surveys for adult medicine specialists, as described in Chapter 2 (Violato et al., 2003; Lockyer & Violato, 2004).

## 3.2 Participants

As of December 2010, 404 adult medicine specialists in Alberta had participated in PAR on two occasions, five years apart. Participation in the program was mandatory on both occasions. The entire sample was used in the current study.

## 3.3 Data Collection

Data were provided by Pivotal Research Inc. For each participant, matched pairs of data (Iteration 1 and Iteration 2) were obtained, each with a unique identifying number known only to Pivotal Research Inc. Each data set contained data from the patient, coworker,

medical colleague and self-assessment surveys. Sociodemographic data, including sex, year of graduation from medical school, location of graduation from medical school (Canadian vs. Non-Canadian), location of practice (urban, rural or regional), and subspecialty within adult medicine were also obtained.

## 3.4 Data Analysis

Confirmatory Factor Analyses (CFA) were performed using IBM SPSS Amos Version 20. All other analyses were performed using IBM SPSS Statistics Version 20. With the exception of participant demographics (which was performed only once), all analyses were performed separately for each of the four surveys (patient, medical colleague, coworker and self).

### *3.4.1 Summary of Statistical Methods*

The following summarizes the statistical methods used to answer our specific research questions:

1. How do coworkers, patients, and medical colleagues rate physicians on various items and attributes? These questions were answered using descriptive statistics (minimum, maximum, mean, standard deviation, kurtosis, skewness).
2. What underlying attributes does each survey actually measure, and are they stable over time? Exploratory Factor Analyses (EFAs) were conducted to identify underlying attributes at Iteration 1. These proposed models were tested at Iteration 2 using CFAs.
3. Are the current surveys reliable? Cronbach's alphas were calculated to determine the internal reliability of each survey, and of each attribute within surveys. Reliability across assessors was determined using generalizability theory.
4. Do scores improve over time, and if yes, can predictors of those changes be identified? Repeated measures multivariate analysis of variances (MANOVAs) and paired t tests were used to determine if there are statistical changes in scores over time. Sequential multiple regressions were used to identify predictors of change over time, after controlling for initial scores at Iteration 1.

5. What is the relationship between self-assessment attribute scores and corresponding medical colleague attribute scores? Pearson's correlations will be calculated for each attribute score.

### *3.4.2 Demographics of Participants*

Participant's sex, country of graduation from medical school (Canadian versus non-Canadian), and location of practice (urban, rural or regional) were expressed as percentages. The range, median and mode year of graduation from medical school were calculated. Subspecialties were reported as a frequency, and as an overall percentage of the total sample.

### *3.4.3 Descriptive Statistics (Research Question 1)*

Descriptive statistics including mean, standard deviation, range, skewness, and kurtosis were calculated for each item, for the overall survey, and for each attribute (as determined by the Exploratory Factor Analysis – described below). The percentage of missing data ("unable to assess" or left blank) was calculated for each item.

### *3.4.4 EFA and CFA (Research Question 2)*

Factor analysis can be used to inform validity of the internal structure of the surveys. Specifically, it is used to determine if the items intended to measure a given construct (in this case, a specific physician attribute) are actually measuring that construct. A primary goal is to explain the most item variance in the fewer number of underlying latent constructs. For example, items intended to measure an aspect of attribute A should correlate highly with other items measuring attribute A; in factor analysis, these items will have high pattern coefficients (or "loadings") onto attribute A and lower pattern coefficients onto unrelated attributes.

Using data from Iteration 1, Exploratory Factor Analyses were performed to identify underlying attributes. For all surveys, sample size was > 300 which is considered good for factor analysis (Tabachnick & Fidell, 2007). Missing data were deleted pairwise.

Suitability of the data for factor analysis was determined using Kaiser's measure of sampling adequacy. A value of > 0.6 indicates the data is appropriate for factor analysis (Tabachnick & Fidell, 2007). Several decisions are required by the researcher during EFA. Different decisions can change results and interpretation, and therefore a detailed description of our decisions is imperative. The following decisions were used for the analyses:

1. *Matrix of Association:* Pearson's product-moment correlation matrix was used. It is appropriate as our variables are intervally scaled (Thompson, 2004).
2. *Extraction Method:* Principal Components Analysis was used, as it is the default method in most statistical packages (Thompson, 2004).
3. *Solution Rotation*: Rotation was used to facilitate factor interpretation. Factor extraction was rotated using Orthogonal Varimax rotation to aid in interpretation. This yielded a simple and interpretable structure for all four surveys.
4. *Factor Retention:* Factors were retained using Kaiser's rule (Eigenvalues > 1). This method of factor retention has been used in previous PAR research where it has consistently yielded interpretable and meaningful factors.
5. *Factor Naming:* Items with pattern coefficients > 0.4 on the rotated pattern matrix were included for interpretation. A coefficient of > 0.4 is considered fair (Tabachnick & Fidell, 2007). Items were allowed to load on more than one factor (ie- some items were "complex" variables). The interpretability of the factors was considered and then named based on the underling attribute represented by the related variables.
6. *Calculations of Attribute Scores:* Crude attribute scores were calculated as the mean score of items that had pattern coefficients of > 0.4 on each attribute. True statistical factor scores, which Tabachnick & Fidell (2007) define as "estimates of the scores subjects would have received on each of the factors had they been measured", were not used in the present study due to complexity.

Confirmatory Factor Analyses (CFA), based on the factor structure models derived using Exploratory Factor Analyses at Iteration 1, were performed using IBM SPSS AMOS

Version 20. CFA allows us to test (or "confirm") our hypothesized interrelationships between measured variables (in this case items on surveys) and latent variables (in this case, attributes) (Violato & Hecker, 2007). How well our proposed model from Iteration 1 explains the data at Iteration 2 was statistically determined using "goodness of fit" estimates.

Missing data were replaced with mean estimates and factors were allowed to covary with each other. Rival models included the independence model (which assumes no relationship between the measured variables) and the saturated model (which by definition has perfect fit). Additionally, a model was run using only "pure" variables (items were only allowed to load on the factor with the highest loading).

The goodness of model fit was determined using model Chi Square ($\chi^2$), normal fit index (NFI), comparative fit index (CFI), and root-mean-square error of approximation (RMSEA). $\chi^2$ tests the difference between the sample covariance matrix and the estimated population covariance matrix. If the model fits the data, $\chi^2$ should be non-significant. However, $\chi^2$ tends to be significant with large sample size (Bentler & Bonett, 1980). NFI and CFI both compare the fit of the proposed model against the independence model. A priori, we considered a value of > 0.95 to indicate good fitting models (Hu & Bentler, 1999). Finally, the RMSEA estimates the lack of fit in a model compared to a saturated model. A priori, we considered a value of 0.06 or less to indicate good fit, and values greater than 0.1 to indicate poor-fit model (Hu & Bentler, 1999, Fan, Thompson & Wang, 1999). The models were not modified after initial fit estimates.

*3.4.5 Reliability Analyses (Research Question 3)*

Reliability refers to the consistency, or reproducibility, of the assessment. Cronbach's alphas were calculated to determine the internal consistencies of the overall surveys, and for the attribute subscales. SPSS does not allow for pairwise deletion when calculating Cronbach's alpha, and listwise deletion led to > 75% of the data being excluded. Therefore, missing values were estimated prior to calculating Cronbach's alpha. The

exception was for the patient survey, where the data file was too large to replace missing values. In this case, Cronbach's alpha was calculated using aggregate, rather than raw, data.

For assessments that depend on human raters, interrater consistency is even more important than internal consistencies of the rating scale (Downing, 2004). In the current study, consistency across raters was estimated using generalizability theory. Because raters were unique to the participant (raters were "nested" within participants), the design was a one-facet nested. This design allowed for the calculation of two variances; the "true" variance of physician participants, and an "error" or "residual" variance. The generalizability coefficient ($Ep^2$) was calculated as:

$$Ep^2 = \frac{\text{Physician (variability component)}}{\text{Physician (variability component)} + \text{Error (variability component)}}$$

*3.4.6 Change Over Time (Research Question 4)*

Repeated measures multivariate Analysis of Variances (MANOVAs) were used to determine if scores were higher at Iteration 2 compared to Iteration 1, using a linear combination of attribute scores. MANOVA is a statistical technique that allows us to test for statistically significant mean differences for a set of dependent variables. In this study, Iteration was used as the independent variable, and attribute scores as the dependent variables. Partial eta squared were calculated to estimate the effect sizes. This value represents the proportion of the variance in the dependent variables that can be explained by the independent variable (Pallent, 2010).

The following assumptions of MANOVA were assessed:

1. Normal distribution of the dependent variables was determined by examining for skewness of the attribute scores.
2. Multivariate outliers were identified using critical Mahalanobis' distances.
3. Multicollinearity was assessed by calculating Pearson's r correlations for the attribute scores.

4. The assumption of homogeneity of Variance-Covariance Matrices was tested using Box's M test for homogeneity of dispersion.

Previous PAR publications have used paired t-tests, rather than MANOVAs, to test for mean differences between iterations (Violato, Lockyer & Fidler, 2008). Therefore, to allow for comparisons with previous PAR research paired t-tests were also calculated to test for statistically significance mean differences in overall survey scores between Iterations. If a difference was found, Cohen's d effect sizes were calculated. The advantage of using a repeated MANOVA, over using a paired t-test for each attribute separately, is a reduced risk of type 1 error. However, it is less powerful in prediction compared to paired t-tests when the dependent variables are correlated as is the case of the current study (Tabachnick & Fidell, 2007).

*3.4.7 What Variables Predict Improvement? (Research Question 4)*

Regression is a statistical technique that allows assessment of relationships between dependent and independent variables (Tabachnick & Fidell, 2007). Typically it is used when the intent of the analysis is determining prediction. One would assume that high (or low) scores at Iteration 1 would predict high (or low scores) at Iteration 2. Moreover, the purpose of this analysis was to identify predictors of *change* rather than *absolute* scores at Iteration 2. For both of these reasons, sequential, rather than standard, multiple regression was used. Scores at Iteration 1 were entered into the model first, followed by other independent variables. This allows us to "control for" scores at Iteration 1.

Sequential multiple regressions were performed to identify which variables could predict overall survey scores at Iteration 2, after controlling for scores at Iteration 1. The dependent variable was overall survey score at Iteration 2. The first independent variable to be entered into the model was overall score at Iteration 1. The following remaining independent variables were entered together into the model as a second block: Familiarity with physician (for the coworker and medical colleague survey); total score at Iteration 1 of the other three surveys; years since graduation; location of graduation;

location of practice; and sex. Nominal variables were first transformed into dichotomous "dummy" variables.

The following assumptions of multiple regression were assessed:

1. Adequacy of sample size was calculated using the following formula:
   N > 105 + (# of independent variables)
2. Multicollinearity was assessed by examining the correlations between independent variables.
3. Outliers with standardized residuals > 3.0 or < -3.0 were identified.
4. Residual scatterplots were examined to ensure they are normally distributed and have a straight line relationship with the predicted dependent variable scores.

*3.4.8 Relationship between Self and Medical Colleague Surveys (Research Question 5)*

The relationship between the self and medical colleague surveys was determined by calculating Pearson's r correlations between corresponding attribute scores.

## 3.5 Ethics Approval

The University of Calgary Conjoint Health Research Ethics Board approved the proposal. (ID number: E-23858).

# Chapter 4: Results

This chapter is divided into five sections. It begins with a description of the adult medicine participants. The next four sections are comprised of results specific to each of the four surveys, and addresses the five research questions. Each survey section begins with descriptive statistics for individual items and attribute scores (research question1). Next, results of the Exploratory and Confirmatory Factor Analysis are presented (research question 2). Evidence to support survey reliability (internal consistencies and generalizability coefficients) follows (research question 3). Finally, score differences between iterations are provided, and predictors of these changes are highlighted (research question 4). The relationship between self and medical colleague attribute scores (research question 5) is addressed in the self survey results section.

## 4.1 Description of Participants

Of the 404 adult medicine specialists who participated, 22.5% were female and 77.5% were male. Most graduated from a Canadian medical school (80.2%). The majority (88.9%) practiced in an urban setting, followed by a regional setting (6.4%) and rural location (4.7%). Year of graduation from medical school ranged between 1952 and 1997. The median year of graduation from medical school was 1983, and the mode was 1978. Table 1 presents the subspecialty distribution of participants. The most common was General Internal Medicine (24%) followed by Cardiology (11.6%).

**Table 1** Subspecialties of Participants

| *Subspecialty* | *n* | *%* |
|---|---|---|
| General Internal Medicine | 97 | 24.0 |
| Cardiology | 47 | 11.6 |
| Neurology | 32 | 7.9 |
| Gastroenterology | 31 | 7.7 |
| Dermatology | 26 | 6.4 |
| Respirology | 23 | 5.7 |
| Radiation Oncology | 21 | 5.2 |
| Rheumatology | 21 | 5.2 |
| Critical Care Medicine | 17 | 4.2 |
| Physical Medicine & Rehabilitation | 17 | 4.2 |
| Endocrinology & Metabolism | 16 | 4.0 |
| Nephrology | 16 | 4.0 |
| Infectious Diseases | 12 | 3.0 |
| Medical Oncology | 11 | 2.7 |
| Hematology | 7 | 1.7 |
| Geriatric Medicine | 6 | 1.5 |
| Clinical Immunology & Allergy | 4 | 1.0 |
| Total | 404 | 100.0 |

## 4.2 Coworker Survey

### *4.2.1 Descriptive Statistics: Survey Items*

The mean number of raters per physician was 7.6 for Iteration 1, and 7.3 at Iteration 2. Descriptive statistics, including minimum and maximum, mean, standard deviation, skewness and kurtosis for each survey item are shown in Table 2. Descriptive statistics for familiarity of the coworker with the physician are also shown. Coworkers were familiar with the physician participants; most respondents knew the physician "well" (4 on the likert scale) or "very well" (5 on the likert scale). Respondents used the full range

of the likert scale, however, the mean rating for all items was above 4 (or "top half"). All items were negatively skewed. The missing rate (either "unable to assess" or left blank) was higher than 15% for five items at Iteration 1 and four items at Iteration 2. This highest missing rate was for the item "responds appropriately in emergency situations" – presumably because this item was not directly observable for all coworkers.

**Table 2a** Item Descriptive Statistics for Coworker Survey, Iteration 1

| Iteration 1 (n = 3001)<br>*Item* | *% Missing* | *Min* | *Max* | *Mean* | *S.D.* | *Skew* | *Kurtosis* |
|---|---|---|---|---|---|---|---|
| Familiarity with Physician | 25.2 | 2 | 5 | 4.27 | .65 | -.46 | -.20 |
| 1. Communicates effectively with patients | 6.4 | 1 | 5 | 4.50 | .70 | -1.21 | .95 |
| 2. Verbally communicates effectively | 1.2 | 2 | 5 | 4.51 | .69 | -1.21 | .76 |
| 3. Effectively communicates in writing | 7.0 | 1 | 5 | 4.49 | .67 | -1.11 | .64 |
| 4. Writes legibly | 2.5 | 1 | 5 | 4.07 | .93 | -.75 | -.01 |
| 5. Is courteous to co-workers | 1.2 | 1 | 5 | 4.54 | .71 | -1.48 | 1.61 |
| 6. Concern for co-worker safety | 12.7 | 1 | 5 | 4.47 | .70 | -1.13 | .60 |
| 7. Respects co-workers | 1.5 | 1 | 5 | 4.54 | .68 | -1.46 | 1.89 |
| 8. Collaborates well with co-workers | 1.8 | 1 | 5 | 4.48 | .72 | -1.27 | 1.18 |
| 9. Shows compassion to patients and their families | 7.4 | 1 | 5 | 4.54 | .68 | -1.36 | 1.20 |
| 10. Separates personal values | 18.8 | 2 | 5 | 4.49 | .65 | -1.01 | .27 |
| 11. Is courteous to patients and their families | 5.3 | 2 | 5 | 4.60 | .64 | -1.40 | 1.10 |
| 12. Allows patients to make informed decisions | 10.5 | 2 | 5 | 4.59 | .61 | -1.30 | .99 |
| 13. Accepts responsibility for patient care | 4.5 | 1 | 5 | 4.62 | .61 | -1.56 | 2.30 |
| 14. Is reasonably accessible to patients | 8.3 | 1 | 5 | 4.30 | .77 | -.79 | -.10 |
| 15. Maintains confidentiality of patients | 8.3 | 1 | 5 | 4.68 | .55 | -1.58 | 2.02 |
| 16. Is accessible for communication about patients | 4.3 | 1 | 5 | 4.46 | .71 | -1.15 | .79 |
| 17. Communicates effectively with families | 13.6 | 2 | 5 | 4.47 | .71 | -1.14 | .62 |
| 18. Accepts responsibility for professional actions | 9.7 | 1 | 5 | 4.60 | .63 | -1.47 | 1.77 |
| 19. Responds in emergency situations | 25.1 | 2 | 5 | 4.62 | .60 | -1.43 | 1.31 |
| 20. Participates effectively as a team member | 2.8 | 2 | 5 | 4.57 | .65 | -1.33 | .95 |
| 21. Facilitates the learning of co-workers | 6.8 | 1 | 5 | 4.48 | .72 | -1.19 | .64 |
| 22. Presents him/herself in a professional manner | 0.2 | 1 | 5 | 4.65 | .60 | -1.70 | 2.58 |
| Overall Coworker Survey Score | | 1.50 | 5 | 4.50 | .51 | -1.15 | .88 |

Note. Some items have been abbreviated. See Appendix A for complete survey.

**Table 2b** Item Descriptive Statistics for Coworker Survey, Iteration 2

| Iteration 2 (n = 2884)<br>*Item* | *% Missing* | *Min* | *Max* | *Mean* | *S.D* | *Skew* | *Kurtosis* |
|---|---|---|---|---|---|---|---|
| Familiarity with Physician | 28.4 | 1 | 5 | 4.33 | .65 | -.55 | 0.23 |
| 1. Communicates effectively with patients | 7.4 | 1 | 5 | 4.58 | .68 | -1.54 | 1.83 |
| 2. Verbally communicates effectively | 1.5 | 1 | 5 | 4.59 | .66 | -1.65 | 2.78 |
| 3. Effectively communicates in writing | 6.6 | 1 | 5 | 4.53 | .69 | -1.46 | 2.10 |
| 4. Writes legibly | 5.1 | 1 | 5 | 4.11 | .95 | -.84 | .14 |
| 5. Is courteous to co-workers | 1.1 | 1 | 5 | 4.60 | .70 | -1.88 | 3.56 |
| 6. Concern for co-worker safety | 11.2 | 1 | 5 | 4.58 | .65 | -1.59 | 2.74 |
| 7. Respects co-workers | 1.2 | 1 | 5 | 4.61 | .66 | -1.81 | 3.49 |
| 8. Collaborates well with co-workers | 1.5 | 1 | 5 | 4.53 | .73 | -1.58 | 2.46 |
| 9. Shows compassion to patients and their families | 7.6 | 1 | 5 | 4.61 | .66 | -1.74 | 2.84 |
| 10. Separates personal values | 15.9 | 1 | 5 | 4.58 | .64 | -1.45 | 1.92 |
| 11. Is courteous to patients and their families | 6.7 | 1 | 5 | 4.65 | .63 | -1.90 | 3.64 |
| 12. Allows patients to make informed decisions | 10 | 1 | 5 | 4.65 | .60 | -1.66 | 2.39 |
| 13. Accepts responsibility for patient care | 4.3 | 1 | 5 | 4.68 | .59 | -1.98 | 4.18 |
| 14. Is reasonably accessible to patients | 8.5 | 1 | 5 | 4.38 | .75 | -.96 | .20 |
| 15. Maintains confidentiality of patients | 6.7 | 1 | 5 | 4.73 | .54 | -2.10 | 4.99 |
| 16. Is accessible for communication about patients | 4.9 | 1 | 5 | 4.55 | .66 | -1.30 | 1.04 |
| 17. Communicates effectively with families | 13.3 | 1 | 5 | 4.54 | .70 | -1.40 | 1.27 |
| 18. Accepts responsibility for professional actions | 8.7 | 2 | 5 | 4.70 | .57 | -1.85 | 3.06 |
| 19. Responds in emergency situations | 22.7 | 1 | 5 | 4.67 | .59 | -1.92 | 4.06 |
| 20. Participates effectively as a team member | 2.1 | 1 | 5 | 4.62 | .66 | -1.81 | 3.37 |
| 21. Facilitates the learning of co-workers | 6.6 | 1 | 5 | 4.58 | .66 | -1.49 | 1.65 |
| 22. Presents him/herself in a professional manner | 0.4 | 1 | 5 | 4.71 | .58 | -2.04 | 3.83 |
| Overall Coworker Survey Score | | 1.73 | 5 | 4.57 | .51 | -1.63 | 2.85 |

Note: Some items have been abbreviated. See Appendix A for complete survey

*4.2.2 Descriptive Statistics: Attribute Scores*

Descriptive statistics for attribute scores (identified by the Exploratory Factor Analysis) are shown in Table 3. Similar to the item and overall survey scores, attribute scores were high and negatively skewed.

**Table 3** Descriptive Statistics for Coworker Attribute Scores

| *Attribute* | *n* | *Min* | *Max* | *Mean* | *S.D.* | *Skewness* | *Kurtosis* |
|---|---|---|---|---|---|---|---|
| **Iteration 1** | | | | | | | |
| Professionalism | 3001 | 1.50 | 5.00 | 4.49 | .51 | -1.10 | .77 |
| Collaborator | 3000 | 1.43 | 5.00 | 4.53 | .55 | -1.29 | 1.27 |
| Psychosocial/Communication | 2993 | 1.86 | 5.00 | 4.51 | .57 | -1.22 | .99 |
| **Iteration 2** | | | | | | | |
| Professionalism | 2884 | 1.75 | 5.00 | 4.56 | .50 | -1.52 | 2.31 |
| Collaborator | 2883 | 1.56 | 5.00 | 4.61 | .54 | -1.82 | 3.89 |
| Psychosocial/Communication | 2872 | 1.00 | 5.00 | 4.59 | .57 | -1.64 | 2.78 |

*4.2.3 Exploratory Factor Analysis*

Principal components analysis was performed on the 22 survey items. The survey was suitable for factor analysis, as the Kaiser measure of sampling adequacy was 0.97. Using Kaiser's Rule (Eigenvalues > 1), three factors were extracted, which explained 65.91% of the variance. The varimax-rotated pattern coefficient matrix is shown in Table 4. The majority of variables loaded on only one factor. Factors were interpretable, in that the items reflected the attributes that they loaded onto. Suggested factor interpretations/labels are: Professionalism (Factor 1), Collaborator (Factor 2), and Psychosocial/Communication (Factor 3).

**Table 4** Varimax-Rotated Pattern Coefficient Matrix, Coworker Survey

| | *Factor* | | |
|---|---|---|---|
| *Item* | 1 | 2 | 3 |
| 1. Communicates effectively with patients | .331 | .272 | **.763** |
| 2. Verbally communicates effectively | .389 | **.520** | **.441** |
| 3. Effectively communicates in writing | **.598** | .237 | .377 |
| 4. Writes legibly | **.456** | .027 | .287 |
| 5. Is courteous to co-workers | .147 | **.778** | .339 |
| 6. Demonstrates appropriate concern for co-worker safety | .337 | **.686** | .222 |
| 7. Respects the professional knowledge and skills of co-workers | .297 | **.787** | .231 |
| 8. Collaborates well with co-workers | .302 | **.765** | .313 |
| 9. Shows compassion to patients and their families | .259 | .332 | **.777** |
| 10. Separates personal values from the management of patients | **.478** | .361 | **.474** |
| 11. Is courteous to patients and their families | .300 | .382 | **.748** |
| 12. Respects the rights of patients to make informed decisions | **.505** | .329 | **.535** |
| 13. Accepts responsibility for patient care | **.698** | .356 | .290 |
| 14. Is reasonably accessible to patients | **.662** | .253 | .276 |
| 15. Maintains confidentiality of patients | **.673** | .323 | .238 |
| 16. Is accessible for appropriate communication about patients | **.652** | .344 | .251 |
| 17. Communicates effectively with families | **.428** | .283 | **.733** |
| 18. Accepts responsibility for professional actions | **.722** | **.401** | .253 |
| 19. Responds appropriately in emergency situations | **.687** | .367 | .225 |
| 20. Participates effectively as a member of the health care team | **.526** | **.599** | .271 |
| 21. Facilitates the learning of co-workers | **.447** | **.559** | .236 |
| 22. This doctor presents him/herself in a professional manner | **.501** | **.488** | .294 |
| Eigenvalue | 12.43 | 1.07 | 1.01 |
| % Variance | 56.49 | 4.86 | 4.57 |

Note. Loadings > 0.4 are in bold

Some items have been abbreviated. See Appendix A for complete survey.

*4.2.4 Confirmatory Factor Analysis (CFA)*

CFA, based on the factor model derived using Exploratory Factor Analysis at Iteration 1, was performed using IBM SPSS AMOS Version 20. The main purpose of this analysis was to test the model factor structure derived at Iteration 1. A rival model, using only pure variables, was also used. For this rival model, items were only allowed to load to one factor (the one with the highest loading). For example survey item 2 loaded both to Factor 2 (loading = 0.520) and to Factor 3 (loading = 0.441). For the initial model, both of these loadings were used. For the rival model, item 2 loaded exclusively to Factor 2. Graphics of the two models are shown in Appendix E. Model fit statistics of the initial and rival models are shown in Table 5.

**Table 5** Model Fit Statistics, Coworker Survey

| *Test* | *Initial Model* | *Rival Model* |
|---|---|---|
| $X^2$ | 7533, df = 225, p=0.000 | 5444, df = 225, p = 0.000 |
| NFI | 0.87 | 0.84 |
| CFI | 0.89 | 0.85 |
| RMSEA | 0.09 | 0.11 |

Note. $X^2$ = Chi Square

NFI = Normalized fit index

CFI = Confirmatory Fit Index

RMSEA = Root Mean Squared Error of Approximation

*4.2.5 Reliability Analyses*

Cronbach's alphas were calculated to determine the internal consistencies of the overall survey, and for the attribute scales. As shown in Table 6, internal consistencies were high at both iterations for the overall survey, and for each attribute score. All Cronbach's alphas were > 0.9, indicating excellent scale reliability.

**Table 6** Cronbach's alphas, Coworker Survey

| | *Iteration 1* | *Iteration 2* |
|---|---|---|
| Overall survey | 0.95 | 0.96 |
| Professionalism | 0.93 | 0.93 |
| Collaborator | 0.92 | 0.93 |
| Psychosocial/Communication | 0.91 | 0.92 |

Consistency across raters was estimated using generalizability theory. Because raters were unique to the participant (raters were "nested" within participants), the design was a one-facet nested. The generalizability coefficients for the overall surveys, and for the attribute scales are shown in Table 7. All but one were > 0.7, indicated acceptable generalizability.

**Table 7** Generalizability Coefficients, Coworker Survey

| | *Iteration 1* | *Iteration 2* |
|---|---|---|
| Overall survey | 0.79 | 0.78 |
| Professionalism | 0.74 | 0.76 |
| Collaborator | 0.76 | 0.69 |
| Psychosocial/Communication | 0.78 | 0.80 |

*4.2.6 Change over Time*

A repeated measures multivariate analysis of variance (MANOVA) was performed to determine if there were differences in scores between Iteration 1 and 2. The hypothesis is that scores will increase from Iteration 1 to Iteration 2. The independent variable was iteration and dependent variables were the three attribute subscores Effect sizes (partial eta) were also calculated to determine the magnitude of change.

Assumption testing showed that the number of cases far exceeded the number of dependent variables. The dependent variables were not normally distributed (Table 3). Using Mahalanobis distances, only seven multivariate outliers were identified at each

iteration. Assumptions of multicollinearity were violated; as shown in Table 8, the dependent variables were highly correlated with each other.

**Table 8** Correlations between Attribute Scores, Coworker Survey

| | Professionalism | Collaboration | Psychosocial/ Communication |
|---|---|---|---|
| Professionalism | 1 | | |
| Collaboration | 0.899 | 1 | |
| Psychosocial/Communication | 0.870 | 0.831 | 1 |

Note. All correlations were significant, 2 tailed, p=0.000.

There was a statistically significant difference between Iteration 1 and Iteration 2 on the combined dependent variables, F (3,387) = 7.26, p = 0.00; Wilks' lambda = 0.95; partial eta squared = 0.053. All dependent variables made significant unique contributions (p = 0.000 for all); Professionalism Score F(1,389) = 20.98; partial eta squared = 0.051, Collaboration Score F (1,389) = 20.10; partial eta squared = 0.051, and Psychosocial/Communication Score F (1,389) = 18.95; partial eta squared = 0.046.

Paired t-tests were also calculated for the overall score and attribute scores. All showed significant improvement over time, but effect sizes were small; Overall Coworker Score $t_{389} = -4.55$, p = 0.000; Cohen's d = 0.24, Professionalism Score $t_{389} = -4.58$, p = 0.000; Cohen's d = 0.25, Collaboration score $t_{389} = -4.48$, p = 0.000, Cohen's d = 0.23, Psychosocial/Communication Score $t_{389} = -4.35$, p = 0.000, Cohen's d = 0.17.

*4.2.7 Predictors of Improvement*

Sequential multiple regression was performed to identify predictors of overall coworker survey score at Iteration 2, after controlling for overall coworker scores at Iteration 1. The following independent variables were used: Coworker familiarity with physician; overall score at Iteration 1 for self, patient, and medical colleague surveys; years since graduation; location of graduation; location of practice; and sex. No major assumptions of multiple regression were violated. Sample size was adequate and multicollinearity was

absent. Additionally, tolerances were all > 0.1 and variation inflation factors were all < 10.

The multiple regression model summary is shown in Table 8. Total variance of overall coworker score at Iteration 2 that is explained by the entire model (including coworker score at Iteration 1) is 33.8%. After controlling for coworker score at Iteration 1, the other variables only contribute 5% of this variance ($F(10,371) = 18.96$, $p = 0.000$). Only two variables made significant unique contributions; overall patient score at Iteration 1 contributed 1.37% of the variance (Beta = 0.12, $p < 0.01$), and overall medical colleague score at Iteration 1 explained 1.17% of the variance (Beta = 0.12, $p < 0.01$).

**Table 9** Model Summary of Sequential Multiple Regression, Coworker Survey

| *Model* | *R* | $R^2$ | *Adjusted* $R^2$ | *S.E.* | $R^2$ *Change* | *F* | *df* | *Sig* |
|---|---|---|---|---|---|---|---|---|
| 1[a] | .537 | .289 | .287 | .259 | .289 | 154.188 | 1,380 | .000 |
| 2[b] | .582 | .338 | .320 | .253 | .050 | 18.957 | 10,371 | .000 |

a. Model 1 Predictors: Overall coworker score at Iteration 1

b. Model 2 Predictors: All independent variables, including overall coworker score at Iteration 1

## 4.3 Medical Colleague Survey

### *4.3.1 Descriptive Statistics: Survey Items*

The mean number of raters per physician was 7.6 at Iteration 1, and 7.4 at Iteration 2. The majority of medical colleagues were peers (50.4% at Iteration 1 and 49.2% at Iteration 2), followed by referring physicians (26.8% Iteration 1 and 25.6% at Iteration 2), and consultants (20.4% at Iteration 1 and 24.0% at Iteration 2). Descriptive statistics for rater familiarity are shown in Table 9. Of those who responded, over 90% knew the physician "well" (4 on the Likert scale) or "very well" (5 on the Likert scale).

Descriptive statistics, including minimum and maximum, mean, standard deviation, skewness and kurtosis for each survey item are shown in Table 9. Respondents used the full range of the Likert scale, however, the mean rating for all items was above 4 (or "top half"). All items were negatively skewed. Many items had missing data rates of > 15% (either left blank or "unable to assess"); 10/38 at Iteration 1 and 8/38 at Iteration 2. The highest missing rate was for the item "makes appropriate use of community resources for psychosocial aspects of care" – presumably because this behavior was not routinely observable by medical colleagues.

**Table 10a** Item Descriptive Statistics for Medical Colleague Survey, Iteration 1

| Iteration 1 (n = 3053) *Item* | *% Missing* | *Min* | *Max* | *Mean* | *S.D.* | *Skew* | *Kurtosis* |
|---|---|---|---|---|---|---|---|
| Familiarity with physician | 23.5 | 1 | 5 | 4.38 | .67 | -.832 | .41 |
| 1. Communicates effectively with patients | 2.8 | 1 | 5 | 4.45 | .67 | -.96 | .37 |
| 2. Communicates effectively with families | 13.0 | 1 | 5 | 4.40 | .69 | -.86 | .15 |
| 3. Communicates with other professionals | 0.9 | 1 | 5 | 4.54 | .63 | -1.16 | .75 |
| 4. Communicates treatment options to patients | 5.2 | 1 | 5 | 4.49 | .63 | -.95 | .43 |
| 5. Performs technical procedures skillfully | 20.3 | 2 | 5 | 4.56 | .60 | -1.02 | .08 |
| 6. Selects diagnostic tests appropriately | 3.9 | 1 | 5 | 4.55 | .59 | -.98 | .28 |
| 7. Critically assesses diagnostic information | 2.4 | 2 | 5 | 4.62 | .56 | -1.15 | .60 |
| 8. Makes the correct diagnosis | 1.1 | 2 | 5 | 4.61 | .56 | -1.11 | .37 |
| 9. Selects appropriate treatments | 1.5 | 2 | 5 | 4.61 | .56 | -1.08 | .30 |
| 10. Maintains quality medical records | 17.0 | 1 | 5 | 4.45 | .67 | -.99 | .50 |
| 11. Handles transfer of care appropriately | 10.5 | 2 | 5 | 4.49 | .65 | -1.00 | .35 |
| 12. Clear about responsibility of continuing care | 6.5 | 2 | 5 | 4.48 | .65 | -.96 | .22 |
| 13. Recognizes psychosocial aspects of illness | 12.0 | 2 | 5 | 4.31 | .70 | -.60 | -.36 |
| 14. Maintains confidentiality | 13.4 | 3 | 5 | 4.54 | .60 | -.92 | -.16 |
| 15. Co-ordinates care effectively | 3.3 | 1 | 5 | 4.51 | .63 | -1.04 | .54 |
| 16. Manages patients with complex problems | 2.9 | 2 | 5 | 4.57 | .60 | -1.14 | .70 |
| 17. Respects the rights of patients | 5.9 | 2 | 5 | 4.51 | .61 | -.86 | -.18 |
| 18. Shows compassion for patients | 5.7 | 2 | 5 | 4.45 | .67 | -.92 | .13 |
| 19. Collaborates with physician colleagues | 1.6 | 1 | 5 | 4.54 | .63 | -1.18 | .97 |
| 20. Is involved with professional development | 18.1 | 1 | 5 | 4.49 | .65 | -1.06 | .58 |
| 21. Accepts responsibility for professional actions | 7.6 | 1 | 5 | 4.54 | .59 | -.97 | .32 |
| 22. Manages health care resources efficiently | 15.9 | 2 | 5 | 4.37 | .66 | -.64 | -.34 |
| 23. Makes appropriate use of resources | 41.1 | 2 | 5 | 4.28 | .70 | -.53 | -.56 |
| 24. Gives priority to urgent requests | 5.2 | 2 | 5 | 4.56 | .62 | -1.11 | .37 |
| 25. Handles emergency situations effectively | 17.4 | 2 | 5 | 4.55 | .61 | -1.08 | .45 |
| 26. Manages own stress effectively | 28.2 | 1 | 5 | 4.26 | .75 | -.68 | -.20 |
| 27. Participates in a system of call | 15.4 | 1 | 5 | 4.42 | .69 | -.97 | .72 |
| 28. Recognizes his/her own limitations | 10.0 | 2 | 5 | 4.38 | .65 | -.62 | -.44 |
| 29. Handles consultation requests timely | 4.7 | 1 | 5 | 4.44 | .67 | -.94 | .33 |
| 30. Advises if referral is outside practice scope | 21.8 | 2 | 5 | 4.45 | .62 | -.71 | -.43 |
| 31. Assumes appropriate responsibility for patients | 1.3 | 2 | 5 | 4.52 | .62 | -1.00 | .33 |
| 32. Information to referring physicians is timely | 6.5 | 1 | 5 | 4.50 | .63 | -1.03 | .71 |
| 33. Critically evaluates the medical literature | 14.4 | 1 | 5 | 4.56 | .61 | -1.10 | .47 |
| 34. Facilitates the learning of others | 8.4 | 1 | 5 | 4.49 | .67 | -1.14 | .93 |
| 35. Contributes to QI and practice guidelines | 28.0 | 1 | 5 | 4.45 | .69 | -1.03 | .39 |
| 36. Participates effectively as a team member | 2.8 | 1 | 5 | 4.53 | .63 | -1.13 | .90 |
| 37. Professional towards physician colleagues | 0.8 | 2 | 5 | 4.60 | .59 | -1.30 | 1.08 |
| 38. I would rate this physician | 0.4 | 2 | 5 | 4.64 | .58 | -1.48 | 1.77 |
| Overall Medical Colleague Survey Score | | 1.50 | 5 | 4.50 | .51 | -1.15 | .88 |

Note: Some items have been abbreviated. See Appendix B for complete survey

**Table 10a** Item Descriptive Statistics for Medical Colleague Survey, Iteration 2

| Iteration 2 (n = 2989) *Item* | *% Missing* | *Min* | *Max* | *Mean* | *S.D.* | *Skew* | *Kurtosis* |
|---|---|---|---|---|---|---|---|
| 1. Communicates effectively with patients | 3.2 | 2 | 5 | 4.55 | .61 | -1.08 | .30 |
| 2. Communicates effectively with patients' families | 10.2 | 2 | 5 | 4.52 | .63 | -1.02 | .23 |
| 3. Communicates effectively with other professionals | 1.2 | 2 | 5 | 4.62 | .60 | -1.41 | 1.38 |
| 4. Communicates treatment options to patients | 4.9 | 2 | 5 | 4.60 | .58 | -1.18 | .74 |
| 5. Performs technical procedures skillfully | 21.8 | 2 | 5 | 4.64 | .56 | -1.36 | 1.10 |
| 6. Selects diagnostic tests appropriately | 3.2 | 1 | 5 | 4.64 | .55 | -1.24 | 1.02 |
| 7. Critically assesses diagnostic information | 2.3 | 2 | 5 | 4.70 | .51 | -1.45 | 1.44 |
| 8. Makes the correct diagnosis following consultation | 1.8 | 2 | 5 | 4.69 | .51 | -1.42 | 1.24 |
| 9. Selects appropriate treatments | 2.0 | 1 | 5 | 4.69 | .52 | -1.52 | 2.00 |
| 10. Maintains quality medical records | 13.5 | 1 | 5 | 4.54 | .65 | -1.34 | 1.73 |
| 11. Handles transfer of care appropriately | 8.7 | 1 | 5 | 4.57 | .63 | -1.32 | 1.45 |
| 12. Clear about responsibility of continuing care | 4.9 | 2 | 5 | 4.58 | .61 | -1.27 | .92 |
| 13. Recognizes psychosocial aspects of illness | 11.2 | 1 | 5 | 4.43 | .68 | -.94 | .38 |
| 14. Maintains confidentiality | 9.9 | 3 | 5 | 4.64 | .55 | -1.22 | .51 |
| 15. Co-ordinates care effectively | 2.8 | 1 | 5 | 4.63 | .58 | -1.42 | 1.73 |
| 16. Manages patients with complex problems | 2.7 | 1 | 5 | 4.64 | .58 | -1.46 | 1.70 |
| 17. Respects the rights of patients | 4.8 | 2 | 5 | 4.63 | .56 | -1.22 | .57 |
| 18. Shows compassion for patients and their families | 4.0 | 1 | 5 | 4.57 | .61 | -1.26 | 1.35 |
| 19. Collaborates with physician colleagues | 1.5 | 2 | 5 | 4.63 | .59 | -1.42 | 1.37 |
| 20. Is involved with professional development | 16.1 | 2 | 5 | 4.60 | .60 | -1.30 | .90 |
| 21. Accepts responsibility for o professional actions | 5.4 | 2 | 5 | 4.64 | .56 | -1.34 | 1.13 |
| 22. Manages health care resources efficiently | 11.4 | 1 | 5 | 4.51 | .63 | -1.00 | .51 |
| 23. Makes appropriate use of community resources | 36.4 | 2 | 5 | 4.44 | .66 | -.85 | -.01 |
| 24. Gives priority to urgent requests | 4.2 | 1 | 5 | 4.65 | .57 | -1.52 | 2.03 |
| 25. Handles emergency situations effectively | 18.7 | 2 | 5 | 4.62 | .58 | -1.34 | 1.10 |
| 26. Manages own stress effectively | 25.9 | 1 | 5 | 4.40 | .70 | -.90 | .27 |
| 27. Participates in a system of call | 16.4 | 1 | 5 | 4.55 | .64 | -1.33 | 1.71 |
| 28. Recognizes his/her own limitations | 7.7 | 2 | 5 | 4.52 | .62 | -.99 | .29 |
| 29. Handles consultation requests in a timely manner | 3.6 | 1 | 5 | 4.53 | .66 | -1.26 | 1.27 |
| 30. Advises if referral is outside the scope of practice | 17.6 | 1 | 5 | 4.58 | .59 | -1.14 | .68 |
| 31. Assumes appropriate responsibility for patients | 0.9 | 1 | 5 | 4.61 | .59 | -1.39 | 1.73 |
| 32. Information to referring physicians is timely | 4.6 | 1 | 5 | 4.62 | .59 | -1.50 | 2.51 |
| 33. Critically evaluates the medical literature | 12 | 1 | 5 | 4.63 | .57 | -1.36 | 1.47 |
| 34. Facilitates the learning of others | 7.9 | 2 | 5 | 4.59 | .62 | -1.27 | .77 |
| 35. Contributes to QI and practice guidelines | 25.2 | 1 | 5 | 4.58 | .63 | -1.38 | 1.63 |
| 36. Participates effectively as a team member | 2.5 | 2 | 5 | 4.62 | .59 | -1.37 | 1.31 |
| 37. Professional towards physician colleagues | 0.6 | 1 | 5 | 4.68 | .57 | -1.77 | 3.39 |
| 38. I would rate this physician | 0.5 | 1 | 5 | 4.70 | .55 | -1.82 | 3.40 |
| Overall Medical Colleague Survey Score | | 2.19 | 5 | 4.59 | .46 | -1.21 | 1.16 |

Note. Some items have been abbreviated. See Appendix B for complete survey.

*4.3.2 Descriptive Statistics: Attribute Scores*

Descriptive statistics for attribute scores (identified by the Exploratory Factor Analysis) are shown in Table 11. Similar to the item and overall survey scores, attribute scores were high and negatively skewed.

**Table 11** Descriptive Statistics for Medical Colleague Attribute Scores

| *Attribute* | *n* | *Min* | *Max* | *Mean* | *S.D.* | *Skewness* | *Kurtosis* |
|---|---|---|---|---|---|---|---|
| **Iteration 1** | | | | | | | |
| Professionalism | 3053 | 2.50 | 5.00 | 4.46 | .51 | -.82 | -.008 |
| Collaboration/Communication | 3053 | 2.17 | 5.00 | 4.50 | .53 | -1.00 | .44 |
| Clinical Competence | 3052 | 1.75 | 5.00 | 4.58 | .49 | -1.17 | 1.10 |
| Professional Development | 3025 | 1.25 | 5.00 | 4.49 | .57 | -1.09 | .94 |
| **Iteration 2** | | | | | | | |
| Professionalism | 2989 | 2.18 | 5.00 | 4.57 | .48 | -1.16 | .94 |
| Collaboration/Communication | 2989 | 2.00 | 5.00 | 4.59 | .49 | -1.30 | 1.37 |
| Clinical Competence | 2988 | 1.75 | 5.00 | 4.65 | .46 | -1.44 | 1.93 |
| Professional Development | 2967 | 2.00 | 5.00 | 4.59 | .53 | -1.26 | 1.08 |

*4.3.3 Exploratory Factor Analysis*

Principal components analysis was performed on the 38 survey items. The survey was suitable for factor analysis, as the Kaiser measure of sampling adequacy was 0.98. Using Kaiser's Rule (Eigenvalues > 1), four factors were extracted, which explained 69.16% of the variance. The varimax-rotated pattern coefficient matrix is shown in Table 12. The majority of variables loaded on only one factor. Factors were interpretable, in that the items reflected the attributes that they loaded onto. Suggested factor labels are: Professionalism (Factor 1), Collaboration/Communication (Factor 2), Clinical Competence (Factor 3), Professional Development (Factor 4).

**Table 12** Varimax-Rotated Pattern Coefficient Matrix, Medical Colleague Survey

| | *Factor* | | | |
|---|---|---|---|---|
| *Item* | 1 | 2 | 3 | 4 |
| 1. Communicates effectively with patients | .201 | **.820** | .288 | .191 |
| 2. Communicates effectively with patients' families | .227 | **.826** | .237 | .201 |
| 3. Communicates effectively with other health care professionals | .330 | **.605** | .374 | .221 |
| 4. Communicates treatment options to patients | .287 | **.684** | .367 | .196 |
| 5. Performs technical procedures skillfully | .393 | .182 | **.637** | .132 |
| 6. Selects diagnostic tests appropriately | .369 | .210 | **.715** | .233 |
| 7. Critically assesses diagnostic information | .301 | .252 | **.742** | .298 |
| 8. Makes the correct diagnosis following consultation | .288 | .306 | **.747** | .247 |
| 9. Selects appropriate treatments | .316 | .303 | **.719** | .266 |
| 10. Maintains quality medical records | **.476** | .369 | .350 | .148 |
| 11. Handles transfer of care appropriately | **.574** | **.418** | .361 | .171 |
| 12. Clear about responsibility of continuing care of patients | **.601** | .376 | .370 | .172 |
| 13. Recognizes psychosocial aspects of illness | **.477** | **.600** | .106 | .251 |
| 14. Maintains confidentiality of patients and their families | **.589** | .390 | .363 | .202 |
| 15. Co-ordinates care effectively with others | **.549** | **.450** | .347 | .212 |
| 16. Manages patients with complex problems | .280 | .371 | **.560** | .319 |
| 17. Respects the rights of patients | **.501** | **.526** | .305 | .245 |
| 18. Shows compassion for patients and their families | **.432** | **.674** | .170 | .222 |
| 19. Collaborates with physician colleagues | **.479** | **.468** | .306 | .341 |
| 20. Is involved with professional development | .305 | .260 | .242 | **.716** |
| 21. Accepts responsibility for own professional actions | **.500** | .378 | .376 | .395 |
| 22. Manages health care resources efficiently | **.566** | .296 | .338 | .346 |
| 23. Makes appropriate use of community resources | **.604** | **.486** | .101 | .273 |
| 24. Gives priority to urgent requests | **.669** | .185 | .347 | .141 |
| 25. Handles emergency situations effectively | **.585** | .222 | **.491** | .195 |
| 26. Manages own stress effectively | **.620** | .242 | .163 | .258 |
| 27. Participates in a system of call | **.617** | .256 | .264 | .302 |
| 28. Recognizes his/her own limitations | **.678** | .317 | .269 | .291 |
| 29. Handles requests for consultation in a timely manner | **.704** | .182 | .265 | .201 |
| 30. Advises if referral is outside the scope of his/her practice | **.686** | .210 | .319 | .297 |
| 31. Assumes appropriate responsibility for patients | **.612** | .352 | .392 | .258 |
| 32. Information to referring physicians is timely | **.601** | .285 | .362 | .265 |
| 33. Critically evaluates the medical literature | .270 | .167 | **.445** | **.653** |
| 34. Facilitates the learning of medical colleagues and co-workers | .266 | .268 | .263 | **.770** |
| 35. Contributes to QI programs and practice guidelines | .296 | .236 | .194 | **.779** |
| 36. Participates effectively as a member of the health care team | **.458** | **.458** | .311 | **.435** |
| 37. Professional towards physician colleagues | **.493** | **.440** | .337 | .309 |
| 38. I would rate this physician for a family member | .313 | **.461** | **.538** | .327 |
| Eigenvalue | 22.39 | 1.47 | 1.26 | 1.18 |
| % Variance | 58.86 | 3.86 | 3.32 | 3.11 |

Note. Loadings > 0.4 are in bold.
Some items have been abbreviated. See Appendix B for complete survey

*4.3.4 Confirmatory Factor Analysis*

CFA, based on the factor model derived using Exploratory Factor Analysis at Iteration 1, was performed using IBM SPSS AMOS Version 20. The main purpose of these analyses was to test the model factor structure derived at Iteration 1. A rival model, using only pure variables, was also used. For this rival model, items were to load to only one factor (the one with the highest loading). For example survey item 11 loaded both to Factor 1 (loading = 0.574) and to Factor 2 (loading = 0.418). For the initial model, both of these loadings were used. For the rival model, item 2 loaded exclusively to Factor 2. Graphics of the two models are shown in Appendix F. Model fit statistics of the initial and rival models are shown in Table 13.

**Table 13** Model Fit Statistics, Medical Colleague Survey

| *Test* | *Initial Solution* | *Pure Solution* |
|---|---|---|
| $\chi^2$ | 16206, df = 693, p = 0.00 | 14593, df = 693, p = 0.00 |
| NFI | 0.83 | 0.85 |
| CFI | 0.84 | 0.86 |
| RMSEA | 0.09 | 0.08 |

Note. $\chi^2$ = Chi Square

NFI = Normalized Fit index

CFI = Confirmatory Fit Index

RMSEA = Root Mean Squared Error of Approximation

*4.3.5 Reliability*

Cronbach's alphas were calculated to determine the internal consistencies of the overall survey, and for the attribute score. As shown in Table 14, internal consistency was high at both iterations for the overall survey, and for each attribute score.

**Table 14** Cronbach's Alphas, Medical Colleague Survey

| | *Iteration 1* | *Iteration 2* |
|---|---|---|
| Overall survey | 0.98 | 0.98 |
| Professionalism | 0.96 | 0.97 |
| Collaborator/Communication | 0.95 | 0.95 |
| Clinical Competence | 0.93 | 0.92 |
| Professional Development | 0.87 | 0.88 |

Consistency across raters was estimated using generalizability theory. Because raters were unique to the participant (raters were "nested" within participants), the design was a one-facet nested. The generalizability coefficients for the overall surveys, and for the attribute scales are shown in Table 15.

**Table 15** Generalizability Coefficients, Medical Colleague Survey

| | *Iteration 1* | *Iteration 2* |
|---|---|---|
| Overall survey | 0.71 | 0.70 |
| Professionalism | 0.69 | 0.68 |
| Collaborator/Communication | 0.72 | 0.72 |
| Clinical Competence | 0.65 | 0.70 |
| Professional Development | 0.71 | 0.72 |

*4.3.6 Change over Time*

A repeated measures multivariate analysis of variance (MANOVA) was performed to determine if there are differences in scores between Iteration 1 and 2. The hypothesis is that scores will increase from Iteration 1 to Iteration 2. The independent variable was iteration and dependent variables were the four attribute scores. Effect sizes (partial eta) were also calculated to determine magnitude of change.

Assumption testing showed that the number of cases far exceeded the number of dependent variables. The dependent variables were not normally distributed (Table 11). Using Mahalanobis distances, only three multivariate outliers were identified at Iteration 1 and eight at Iteration 2. Assumptions of multicollinearity were violated; as shown in Table 8, the dependent variables were highly correlated with each other.

**Table 16** Correlations between Attribute Scores, Medical Colleague Survey

| | Professionalism | Collaboration/ Communication | Clinical Competency | Professional Development |
|---|---|---|---|---|
| Professionalism | 1 | | | |
| Collaboration/Communication | 0.945 | 1 | | |
| Clinical Competency | 0.862 | 0.844 | 1 | |
| Professional Development | 0.816 | 0.804 | 0.802 | 1 |

Note. All correlations were significant, 2 tailed, p = 0.000.

There was a statistically significant difference between Iteration 1 and Iteration 2 on the combined dependent variables, F(4,398) = 20.09, p = 0.000; Wilks' lambda = 0.832, partial eta squared = 0.168. All dependent variables made unique contributions (p = 0.000 for all); Professionalism Score F(1,401) = 78.77; partial eta = 0.16, Collaboration/Communication Score F(1,401) = 68.87; partial eta = 0.15, Clinical Competence Score F(1,401) = 49.66; partial eta = 0.11, and Professional Development Score F(1,401) = 60.81; partial eta = 0.16.

Paired t-tests were also used to compare overall and attribute scores between iterations. All showed significant improvement over time; Overall Medical Colleague Score $t_{401}$ = -8.82, p = 0.000; Cohen's d = 0.40, Professionalism Score $t_{401}$ = -8.88, p = 0.000; Cohen's d = 0.44, Collaboration/Communication Score $t_{401}$ = -8.30, p = 0.000, Cohen's d = 0.37, Clinical Competency Score $t_{401}$ = -7.05, p = 0.000, Cohen's d = 0.28, Professional Development Score $t_{401}$ = -7.80, p = 0.000, Cohen's d = 0.35.

*4.3.7 What Variables Predict Improvement over Iterations?*

Sequential multiple regression was performed to identify predictors of overall medical colleague scores at Iteration 2, after controlling for scores at Iteration 1. The following independent variables were used: colleague familiarity with physician; total score at Iteration 1 for self, patient, and coworker surveys; years since graduation; location of graduation; location of practice; and sex. No major assumptions of multiple regression were violated. Sample sizes were adequate and multicollinearity was absent. Tolerances were all > 0.1 and variation inflation factors were all < 10.

The multiple regression model summary is shown in Table 17. Total variance of overall medical colleague score at Iteration 2 that is explained by the entire model (including coworker score at Iteration 1) is 41.7%. However, after controlling for overall medical colleague score at Iteration 1, the other independent variables only contributed 6.8% of this variance ($F(374,10) = 26.74$, $p = 0.000$). None of the other variables made individual unique contributions to the variance.

**Table 17** Model Summary of Sequential Multiple Regression, Medical Colleague Survey

| *Model* | *R* | $R^2$ | *S.E.* | $R^2$ *Change* | *F* | *df* | *Sig* |
|---|---|---|---|---|---|---|---|
| 1[a] | .590 | .349 | .192 | .349 | 205.04 | 1,383 | .000 |
| 2[b] | .646 | .417 | .184 | .068 | 26.74 | 10,374 | .000 |

a. Model Predictors: Overall medical colleague score at iteration 1

b. Model Predictors: All independent variables, including overall medical colleague score at iteration 1

## 4.4 Patient Survey

### *4.4.1 Descriptive Statistics: Survey Items*

The mean number of raters per physician was 23.8 in iteration 1, and 23.6 in Iteration 2. Rater characteristics are given in Table 18. There was a slight predominance of female raters. Most were patients, rather than caregivers of patients. All age groups were represented.

**Table 18** Characteristics of Patient Raters

| | *Iteration 1 (%)* | *Iteration 2 (%)* |
|---|---|---|
| Female/Male | 55.5/44.5 | 53.9/46.1 |
| Patient/Caregiver | 93.9/6.1 | 94.0/6 |
| Age 19-45 | 27.8 | 30.6 |
| 46-65 | 40.3 | 43.6 |
| 66+ | 29.6 | 24.5 |

Descriptive statistics, including minimum and maximum, mean, standard deviation, skewness and kurtosis for each survey item are shown in Table 19. Respondents used the full range of the Likert scale. The mean scores for all items were high, ranging from 4.26 to 4.78, and were negatively skewed. Kurtosis was positive and higher than the other surveys. Additionally, this survey had the highest missing rates (left blank or "unable to assess"); 15/40 items at Iteration 1 and 14/40 items at Iteration 2 had missing rates over 15%.

**Table 19a** Item Descriptive Statistics for Patient Survey, Iteration 1

| Iteration 1 (n = 9354) *Item* | % Missing | Min | Max | Mean | S.D. | Skew | Kurtosis |
|---|---|---|---|---|---|---|---|
| 1. Explained illness or concern to me clearly | 2.4 | 1 | 5 | 4.67 | .61 | -2.65 | 10.54 |
| 2. Explained treatment choices/options | 5.2 | 1 | 5 | 4.64 | .63 | -2.38 | 8.52 |
| 3. Explained my follow-up plan | 6.4 | 1 | 5 | 4.62 | .65 | -2.33 | 7.86 |
| 4. Told me how and when to take my medicine | 25.0 | 1 | 5 | 4.63 | .64 | -2.36 | 8.13 |
| 5. Told me the side effects of the medicine | 28.3 | 1 | 5 | 4.46 | .80 | -1.71 | 3.37 |
| 6. Spends enough time with me | 1.4 | 1 | 5 | 4.62 | .66 | -2.38 | 8.00 |
| 7. Shows interest in my problems | 1.3 | 1 | 5 | 4.67 | .62 | -2.62 | 9.93 |
| 8. Asks details about my personal life | 8.8 | 1 | 5 | 4.43 | .77 | -1.59 | 3.26 |
| 9. Answers my questions well | 2.1 | 1 | 5 | 4.66 | .62 | -2.53 | 9.47 |
| 10. Examines me appropriately for my problems | 3.1 | 1 | 5 | 4.66 | .61 | -2.50 | 9.53 |
| 11. Treats me with respect | 1.3 | 1 | 5 | 4.75 | .57 | -3.34 | 15.72 |
| 12. Helps me with my worries and fears | 7.6 | 1 | 5 | 4.55 | .71 | -1.95 | 5.09 |
| 13. Office is easy to get into | 8.3 | 1 | 5 | 4.36 | .84 | -1.64 | 3.21 |
| 14. Office has appropriate waiting areas | 2.9 | 1 | 5 | 4.53 | .67 | -1.96 | 6.35 |
| 15. Examining rooms are adequately | 3.8 | 1 | 5 | 4.56 | .66 | -2.07 | 6.98 |
| 16. Office is clean and in good repair | 2.8 | 1 | 5 | 4.60 | .63 | -2.23 | 8.40 |
| 17. Office provides adequate privacy | 2.9 | 1 | 5 | 4.61 | .63 | -2.31 | 8.75 |
| 18. I can reach the office by phone during the day | 13.4 | 1 | 5 | 4.45 | .72 | -1.70 | 4.52 |
| 19. Received explanation if appointment delayed | 31.6 | 1 | 5 | 4.37 | .77 | -1.42 | 2.80 |
| 20. My messages are returned | 28.6 | 1 | 5 | 4.46 | .74 | -1.68 | 4.00 |
| 21. The staff are helpful and pleasant | 3.0 | 1 | 5 | 4.64 | .60 | -2.40 | 9.53 |
| 22. The staff are respectful of patients | 3.4 | 1 | 5 | 4.65 | .60 | -2.47 | 10.05 |
| 23. The staff behave in a professional manner | 3.5 | 1 | 5 | 4.66 | .59 | -2.52 | 10.52 |
| 24. The staff work well with the doctor | 13.1 | 1 | 5 | 4.64 | .61 | -2.32 | 8.77 |
| 25. The staff ensures confidentiality | 16.7 | 1 | 5 | 4.55 | .69 | -1.96 | 5.62 |
| 26. In an emergency, office provides instructions | 46.4 | 1 | 5 | 4.41 | .79 | -1.51 | 2.74 |
| 27. This doctor provides reports to my family doctor | 18.8 | 1 | 5 | 4.60 | .65 | -2.18 | 7.18 |
| 28. Provides insurance/legal reports | 57.2 | 1 | 5 | 4.41 | .79 | -1.42 | 2.32 |
| 29. Provides reports, files, or copies of letters | 43.5 | 1 | 5 | 4.52 | .71 | -1.77 | 4.42 |
| 30. Arranges appointments with other specialists | 36.4 | 1 | 5 | 4.60 | .66 | -2.09 | 6.34 |
| 31. Office follows-up on serious problems | 32.1 | 1 | 5 | 4.62 | .66 | -2.31 | 7.72 |
| 32. I am told what to do if problem does improve | 20.5 | 1 | 5 | 4.58 | .68 | -2.17 | 6.88 |
| 33. Asked about non-prescription medicine | 6.6 | 1 | 5 | 4.61 | .65 | -2.34 | 8.17 |
| 34. This doctor talks to me about preventative care | 20.6 | 1 | 5 | 4.48 | .75 | -1.73 | 3.75 |
| 35. This doctor has good written health information | 24.8 | 1 | 5 | 4.47 | .73 | -1.64 | 3.68 |
| 36. This doctor refers me to appropriate resources | 36.9 | 1 | 5 | 4.26 | .90 | -1.16 | 1.04 |
| 37. I would go back to this doctor | 3.0 | 1 | 5 | 4.75 | .58 | -3.36 | 15.68 |
| 38. I would send a friend to this doctor | 4.2 | 1 | 5 | 4.74 | .60 | -3.25 | 14.23 |
| 39. This doctor presents in a professional manner | 2.6 | 1 | 5 | 4.78 | .53 | -3.62 | 19.03 |
| 40. I was helped by this doctor | 5.6 | 1 | 5 | 4.74 | .58 | -3.20 | 14.14 |
| Overall Patient Survey Score | | 1 | 5 | 4.58 | .51 | -2.92 | 15.49 |

Note. Some items have been abbreviated. See Appendix C for complete survey.

**Table 19b** Item Descriptive Statistics for Patient Survey, Iteration 2

| Iteration 2 (n = 9247)<br>*Item* | *% Missing* | *Min* | *Max* | *Mean* | *S.D.* | *Skew* | *Kurtosis* |
|---|---|---|---|---|---|---|---|
| 1. Explained illness or concern to me clearly | 3.2 | 1 | 5 | 4.72 | .58 | -2.94 | 12.60 |
| 2. Explained treatment choices/options | 5.4 | 1 | 5 | 4.68 | .61 | -2.67 | 10.30 |
| 3. Explained my follow-up plan | 6.5 | 1 | 5 | 4.66 | .63 | -2.52 | 8.79 |
| 4. Told me how and when to take my medicine | 25.1 | 1 | 5 | 4.68 | .63 | -2.62 | 9.43 |
| 5. Told me the side effects of the medicine | 27.6 | 1 | 5 | 4.51 | .78 | -1.85 | 3.81 |
| 6. Spends enough time with me | 2.2 | 1 | 5 | 4.67 | .64 | -2.67 | 9.72 |
| 7. Shows interest in my problems | 2.1 | 1 | 5 | 4.72 | .61 | -3.01 | 12.33 |
| 8. Asks details about my personal life | 8.5 | 1 | 5 | 4.51 | .74 | -1.80 | 4.09 |
| 9. Answers my questions well | 2.6 | 1 | 5 | 4.70 | .60 | -2.85 | 11.57 |
| 10. Examines me appropriately for my problems | 3.6 | 1 | 5 | 4.70 | .60 | -2.84 | 11.72 |
| 11. Treats me with respect | 2.2 | 1 | 5 | 4.79 | .54 | -3.82 | 20.06 |
| 12. Helps me with my worries and fears | 8.3 | 1 | 5 | 4.61 | .69 | -2.21 | 6.35 |
| 13. Office is easy to get into | 7.0 | 1 | 5 | 4.36 | .88 | -1.62 | 2.71 |
| 14. Office has appropriate waiting areas | 2.7 | 1 | 5 | 4.57 | .66 | -2.02 | 6.29 |
| 15. Examining rooms are adequate | 3.6 | 1 | 5 | 4.61 | .63 | -2.21 | 7.61 |
| 16. Office is clean and in good repair | 2.5 | 1 | 5 | 4.64 | .62 | -2.36 | 8.77 |
| 17. Office provides adequate privacy | 2.7 | 1 | 5 | 4.66 | .61 | -2.48 | 9.54 |
| 18. I can reach the office by phone during the day | 11.9 | 1 | 5 | 4.47 | .74 | -1.77 | 4.28 |
| 19. Received explanation if appointment is delayed | 28.3 | 1 | 5 | 4.43 | .77 | -1.56 | 2.95 |
| 20. My messages are returned | 24.3 | 1 | 5 | 4.51 | .74 | -1.85 | 4.44 |
| 21. The staff are helpful and pleasant | 2.8 | 1 | 5 | 4.69 | .59 | -2.69 | 11.09 |
| 22. The staff are respectful of patients | 2.9 | 1 | 5 | 4.70 | .57 | -2.75 | 11.72 |
| 23. The staff behave in a professional manner | 2.8 | 1 | 5 | 4.71 | .57 | -2.84 | 12.37 |
| 24. The staff work well with the doctor | 10.1 | 1 | 5 | 4.69 | .59 | -2.66 | 10.78 |
| 25. The staff ensure confidentiality | 12.8 | 1 | 5 | 4.60 | .69 | -2.23 | 6.67 |
| 26. In emergency, office provides with instructions | 42.9 | 1 | 5 | 4.45 | .78 | -1.57 | 2.79 |
| 27. This doctor provides reports to my family doctor | 17.2 | 1 | 5 | 4.63 | .65 | -2.37 | 8.03 |
| 28. Provides insurance and medicolegal reports | 53.5 | 1 | 5 | 4.46 | .77 | -1.52 | 2.63 |
| 29. Provides reports, files, or copies of letters | 39.5 | 1 | 5 | 4.57 | .69 | -2.01 | 5.49 |
| 30. Arranges appointments with other specialists | 32.9 | 1 | 5 | 4.64 | .64 | -2.30 | 7.27 |
| 31. Office follows-up on serious problems | 28.5 | 1 | 5 | 4.67 | .63 | -2.53 | 8.94 |
| 32. Told what to do if my problem does not improve | 19.5 | 1 | 5 | 4.62 | .66 | -2.24 | 6.91 |
| 33. Asked about non-prescription medicine | 6.2 | 1 | 5 | 4.66 | .62 | -2.54 | 9.25 |
| 34. This doctor talks to me about preventative care | 18.2 | 1 | 5 | 4.53 | .73 | -1.84 | 4.20 |
| 35. This doctor has good written health information | 22.6 | 1 | 5 | 4.54 | .71 | -1.78 | 4.10 |
| 36. This doctor refers me to appropriate resources | 31.8 | 1 | 5 | 4.37 | .85 | -1.34 | 1.54 |
| 37. I would go back to this doctor | 2.6 | 1 | 5 | 4.79 | .54 | -3.64 | 18.05 |
| 38. I would send a friend to this doctor | 3.6 | 1 | 5 | 4.77 | .57 | -3.51 | 16.12 |
| 39. This doctor presents in a professional manner | 2.4 | 1 | 5 | 4.81 | .51 | -4.04 | 22.78 |
| 40. I was helped by this doctor | 4.4 | 1 | 5 | 4.78 | .55 | -3.60 | 17.64 |
| Overall Patient Survey Score | | 1 | 5 | 4.63 | .50 | -3.20 | 17.53 |

Note. Some items have been abbreviated. See Appendix C for complete survey.

*4.4.2 Descriptive Statistics: Attribute Scores*

Descriptive statistics for attribute scores (identified by the Exploratory Factor Analysis) are shown in Table 20. Similar to the item and overall survey scores, attribute scores were high, negatively skewed, and had a large positive kurtosis.

**Table 20** Descriptive Statistics for Patient Attribute Scores

| *Attribute* | *n* | *Min* | *Max* | *Mean* | *S.D.* | *Skewness* | *Kurtosis* |
|---|---|---|---|---|---|---|---|
| **Iteration 1** | | | | | | | |
| Psychosocial | 9347 | 1.00 | 5.00 | 4.63 | .53 | -3.03 | 15.11 |
| Communication | 9337 | 1.00 | 5.00 | 4.49 | .57 | -2.06 | 8.37 |
| Clinic Staff | 9222 | 1.00 | 5.00 | 4.65 | .52 | -3.11 | 16.75 |
| Office Staff | 9154 | 1.00 | 5.00 | 4.53 | .60 | -2.09 | 8.32 |
| **Iteration 2** | | | | | | | |
| Psychosocial | 9240 | 1.00 | 5.00 | 4.67 | .51 | -3.37 | 17.89 |
| Communication | 9231 | 1.00 | 5.00 | 4.54 | .56 | -2.26 | 9.32 |
| Clinic Staff | 9134 | 1.00 | 5.00 | 4.69 | .50 | -3.37 | 18.49 |
| Office Staff | 9064 | 1.00 | 5.00 | 4.57 | .59 | -2.13 | 8.09 |

*4.4.3 Exploratory Factor Analysis*

Principal component analysis was performed on the 40 survey items. The survey was suitable for factor analysis, as the Kaiser measure of sampling adequacy was 0.98. Using Kaiser's Rule (Eigenvalues > 1), four factors were extracted, which explained 72.34% of the variance. The varimax-rotated pattern coefficient matrix is shown in Table 21. The majority of variables loaded on only one factor. Suggested factor labels are: Psychosocial/Humanistic (Factor 1), Communication (Factor 2), Office Staff (Factor 3), Clinic Space (Factor 4).

**Table 21** Varimax-Rotated Pattern Coefficient Matrix, Patient Survey

| *Item* | *Factor* 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1. Explained illness or concern to me clearly | **.728** | .271 | .200 | .237 |
| 2. Explained treatment choices/options | **.719** | .306 | .174 | .232 |
| 3. Explained my follow-up plan | **.691** | .327 | .172 | .239 |
| 4. Told me how and when to take my medicine | **.638** | .369 | .172 | .278 |
| 5. Told me the side effects of the medicine | **.515** | **.486** | .050 | .206 |
| 6. Spends enough time with me | **.708** | .321 | .182 | .264 |
| 7. Shows interest in my problems | **.761** | .286 | .212 | .263 |
| 8. Asks details about my personal life, when appropriate | **.561** | **.468** | .039 | .241 |
| 9. Answers my questions well | **.754** | .313 | .204 | .265 |
| 10. Examines me appropriately for my problems | **.716** | .303 | .218 | .297 |
| 11. Treats me with respect | **.761** | .203 | .308 | .286 |
| 12. Helps me with my worries and fears | **.665** | **.405** | .119 | .234 |
| 13. Office is easy to get into (wheelchair accessible, parking) | .176 | .271 | .144 | **.673** |
| 14. Office has appropriate waiting areas | .298 | .262 | .264 | **.753** |
| 15. Examining rooms are adequately sized, adequate equipment | .335 | .253 | .289 | **.744** |
| 16. Office is clean and in good repair | .371 | .222 | .342 | **.725** |
| 17. Office provides adequate privacy | .390 | .232 | .338 | **.703** |
| 18. I can reach the office by phone during the day | .189 | **.534** | **.404** | .336 |
| 19. I receive an appropriate explanation if my appointment is delayed | .181 | **.655** | .347 | .297 |
| 20. My messages are returned | .194 | **.605** | **.417** | .298 |
| 21. The staff are helpful and pleasant | .283 | .293 | **.784** | .283 |
| 22. The staff are respectful of patients | .309 | .285 | **.788** | .295 |
| 23. The staff behave in a professional manner | .327 | .284 | **.778** | .300 |
| 24. The staff work well with the doctor | .333 | .353 | **.715** | .281 |
| 25. The staff prevent patients from hearing confidential information | .249 | **.407** | **.592** | .297 |
| 26. In an emergency this office provides me with clear instructions | .290 | **.686** | .268 | .241 |
| 27. This doctor provides reports to my family doctor | **.405** | **.544** | .348 | .249 |
| 28. Provides insurance and medicolegal reports in a timely manner | .329 | **.752** | .249 | .231 |
| 29. Provides reports, files, or copies of letters in a timely manner | .351 | **.707** | .318 | .218 |
| 30. Arranges appointments with other specialists when necessary | **.406** | **.616** | .352 | .180 |
| 31. This doctor's office follows-up on serious problems | **.457** | **.592** | .369 | .171 |
| 32. I am told what to do if my problem does not get better | **.495** | **.589** | .305 | .152 |
| 33. I am asked about prescription and non-prescription medicine | **.479** | **.468** | .322 | .215 |
| 34. This doctor to me about preventative care | **.415** | **.626** | .167 | .135 |
| 35. This doctor has good written health information | .399 | **.652** | .221 | .207 |
| 36. This doctor refers me to appropriate educational resources | .316 | **.721** | .078 | .119 |
| 37. I would go back to this doctor | **.737** | .276 | **.426** | .114 |
| 38. I would send a friend to this doctor | **.730** | .286 | **.414** | .097 |
| 39. This doctor presents him/herself in a professional manner | **.715** | .227 | **.469** | .162 |
| 40. I was helped by this doctor | **.701** | .282 | **.414** | .123 |
| Eigenvalue | 23.81 | 2.14 | 1.71 | 1.28 |
| % Variance | 59.93 | 5.35 | 4.27 | 3.19 |

Note. Loadings > 0.4 are in bold

Some items are abbreviated. See Appendix C for complete survey.

### 4.4.4 *Confirmatory Factor Analysis (CFA)*

CFA, based on the factor model derived using Exploratory Factor Analysiss at Iteration 1, was performed using IBM SPSS AMOS Version 20. The main purpose of these analyses was to test the model factor structure derived at Iteration 1. A rival model, using only pure variables, was also used. For this rival model, items were only allowed to load to one factor (the one with the highest loading). For example, survey item 5 loaded both to Factor 1 (loading = 0.515) and to Factor 2 (loading = 0.486). For the initial model, both of these loadings were used. For the rival model, item 5 loaded exclusively to factor 1. Graphics of the two models are shown in Appendix G. Model fit statistics of the initial and rival models are shown in Table 22.

**Table 22** Model Fit Statistics, Patient Survey

| *Test* | *Initial Model* | *Rival Model* |
|---|---|---|
| $\chi^2$ | 69352, df = 770, p = 0.000 | 48371, df = 770, p = 0.000 |
| NFI | 0.81 | 0.86 |
| CFI | 0.81 | 0.87 |
| RMSEA | 0.10 | 0.08 |

Note. $\chi^2$ = Chi Square

NFI = Normalized fit index

CFI = Confirmatory Fit Index

RMSEA = Root Mean Squared Error of Approximation

### *4.4.5 Reliability*

Cronbach's alphas were calculated to determine the internal consistencies of the overall surveys, and for the attribute subscales. As shown in Table 23, internal consistencies were high at both Iterations for the overall survey, and for each attribute score.

**Table 23** Cronbach's Alphas, Patient Survey

| | *Iteration 1* | *Iteration 2* |
|---|---|---|
| Total survey | 0.98 | 0.98 |
| Psychosocial/Humanistic | 0.98 | 0.98 |
| Communication | 0.96 | 0.97 |
| Office Staff | 0.93 | 0.93 |
| Clinic Space | 0.91 | 0.94 |

Consistency across raters was estimated using generalizability theory. Because raters were unique to the participant (raters were "nested" within participants), the design was a one-facet nested. The generalizability coefficients for the overall surveys, and for the attribute subscales are shown in Table 24

**Table 24** Generalizability Coefficients, Patient Survey

| | *Iteration 1* | *Iteration 2* |
|---|---|---|
| Overall survey | 0.68 | 0.71 |
| Psychosocial/Humanistic | 0.70 | 0.73 |
| Communication | 0.70 | 0.66 |
| Office Staff | 0.65 | 0.69 |
| Clinic Space | 0.60 | 0.70 |

*4.4.6 Change over Time*

A repeated measures multivariate analysis of variance (MANOVA) was performed to determine if there are differences in scores between Iteration 1 and 2. The hypothesis is that scores will increase from Iteration 1 to Iteration 2. The independent variable was iteration and dependent variables were the four attribute scores. Effect sizes (partial eta) were also calculated to determine magnitude of change.

Assumption testing showed that the number of cases far exceeded the number of dependent variables. However, the dependent variables (attribute scores) were not

normally distributed (see Table 20). Using Mahalanobis distances, only 2 multivariate outliers were identified at Iteration 1 and only 1 at Iteration 2. Assumptions of multicollinearity were violated; as shown in Table 25, the dependent variables were highly correlated with each other.

**Table 25** Correlations between Attribute Scores, Patient Survey

| | Psychosocial/ Humanistic | Communication | Office Space | Clinic Space |
|---|---|---|---|---|
| Psychosocial/Humanistic | 1 | | | |
| Communication | 0.93 | 1 | | |
| Office Staff | 0.86 | 0.88 | 1 | |
| Clinic Space | 0.67 | 0.69 | 0.72 | 1 |

Note. All correlations are significant, 2 tailed, p = 0.000.

There was a statistically significant difference between Iteration 1 and Iteration 2 on the combined dependent variables, F(4,381) = 6.96, p = 0.00; Wilks' lambda = 0.932, partial eta squared = 0.068. Attributes 1-3 made unique contributions; Psychosocial Score F(1,384) = 25.13, p = 0.000, partial eta squared = 0.061; Communication Score F(1,384) = 20.93, p = 0.000, partial eta squared = 0.052, and Office Staff Score F(1,384) = 19.11, p = 0.000, partial eta squared = 0.047.

Paired t-tests were also used to determine changes in overall score and attribute scores between iterations. All showed significant improvement over time; Overall Patient Score $t_{385}$ = -4.81, p = 0.000; Cohen's d = 0.26, Psychosocial/Humanistic Score $t_{385}$ = -5.01, p = 0.000; Cohen's d = 0.24, Communication Score $t_{385}$ = -4.58, p = 0.000, Cohen's d = 0.29, Office Staff Score $t_{385}$= -4.37, p = 0.000, Cohen's d = 0.34, Clinic Space Score $t_{385}$ = -2.00, p = 0.046, Cohen's d = 0.15.

*4.4.7 What Variables Predict Improvement over Iterations?*

Sequential multiple regression was performed to identify predictors of overall patient scores at Iteration 2, after controlling for scores at Iteration 1. The following independent variables were used: overall score at Iteration 1 for self, medical colleague, and coworker surveys; years since graduation; location of graduation; location of practice; and sex. No major assumptions of multiple regression were violated. Sample sizes were adequate and multicollinearity was absent. Tolerances were all > 0.1 and variation inflation factors were all < 10.

The multiple regression model summary is shown in Table 26. Total variance of overall patient colleague score at Iteration 2 that is explained by the entire model (including patient score at Iteration 1) is 16.8%. However, after controlling for overall patient score at Iteration 1, the other independent variables only contributed 6.3% of this variance ($F(372,9) = 8.38$, $p = 0.000$). Only two variables made significant unique contributions; overall Medical Colleague Score at Iteration 1 contributed 1.5% of the variance (Beta = 0.134, $p = 0.009$) and overall Self Score at Iteration 1 contributed 1.5% of the variance (Beta = 0.124, $p = 0.01$).

**Table 26** Model Summary of Sequential Multiple Regression, Patient Survey

| *Model* | *R* | $R^2$ | *Adjusted* $R^2$ | *S.E.* | $R^2$ *Change* | *F* | *df* | *Sig* |
|---|---|---|---|---|---|---|---|---|
| 1[a] | 0.325 | .106 | 0.103 | .148 | .106 | 44.90 | 1,380 | .000 |
| 2[b] | 0.410 | .168 | 0.148 | .144 | .063 | 8.38 | 9,372 | .001 |

a. Model Predictors: Overall patient score at iteration 1

b. Model Predictors: All independent variables, including overall patient score at iteration 1.

## 4.5 Self Survey

### *4.5.1 Descriptive Statistics: Survey Items*

Descriptive statistics, including minimum and maximum, mean, standard deviation, skewness and kurtosis for each item are shown in Table 27. Unlike the previous surveys, respondents rarely selected 1 on the Likert scale. It is the only survey in which the mean item scores were normally distributed. All mean item scores were lower than for the medical colleague survey, and ranged from 3.51 (item "I manage my stress effectively") to 4.21 (item "I communicate effectively with patients"). Unlike the previous surveys, no items had missing rates > 15%.

**Table 27a** Item Descriptive Statistics for Self Survey, Iteration 1

| Iteration 1 (n = 401) *Item* | % Missing | Min | Max | Mean | S.D. | Skew | Kurtosis |
|---|---|---|---|---|---|---|---|
| 1. I communicate effectively with patients | 0 | 3 | 5 | 4.21 | .65 | -.24 | -.72 |
| 2. I communicate effectively with patients' families | 1.0 | 2 | 5 | 4.03 | .67 | -.09 | -.59 |
| 3. I communicate effectively with others | 0 | 3 | 5 | 3.99 | .60 | .00 | -.22 |
| 4. I communicate treatment options to patients | 0.5 | 3 | 5 | 4.15 | .64 | -.14 | -.59 |
| 5. I perform technical procedures skillfully | 10.6 | 2 | 5 | 3.98 | .73 | -.01 | -1.00 |
| 6. I select diagnostic tests appropriately | 0 | 3 | 5 | 3.95 | .66 | .06 | -.68 |
| 7. I critically assess diagnostic information | 0.2 | 3 | 5 | 4.02 | .65 | -.02 | -.64 |
| 8. I make the correct diagnosis following consultation | 0.5 | 2 | 5 | 4.02 | .67 | -.07 | -.57 |
| 9. I select appropriate treatments | 0.5 | 3 | 5 | 4.03 | .69 | -.05 | -.90 |
| 10. I maintain quality medical records | 0.2 | 2 | 5 | 3.88 | .75 | -.01 | -.79 |
| 11. I handle transfer of care appropriately | 2.5 | 3 | 5 | 3.80 | .67 | .26 | -.81 |
| 12. Clear about responsibility of continuing care | 0.2 | 2 | 5 | 3.82 | .67 | .08 | -.52 |
| 13. I recognize psychosocial aspects of illness | 0.5 | 2 | 5 | 3.83 | .72 | .02 | -.58 |
| 14. I maintain confidentiality of | 1.0 | 2 | 5 | 4.04 | .72 | -.10 | -.95 |
| 15. I co-ordinate care effectively with others | 0.5 | 2 | 5 | 3.89 | .65 | .06 | -.54 |
| 16. I manage patients with complex problems | 1.5 | 2 | 5 | 4.02 | .69 | -.07 | -.72 |
| 17. I respect the rights of patients | 0 | 3 | 5 | 4.07 | .69 | -.10 | -.88 |
| 18. I show compassion for patients and their families | 0.2 | 3 | 5 | 4.08 | .69 | -.11 | -.91 |
| 19. I collaborate with physician colleagues | 0.2 | 2 | 5 | 3.91 | .67 | .00 | -.49 |
| 20. I am involved with professional development | 1.7 | 2 | 5 | 3.88 | .73 | -.02 | -.65 |
| 21. I accept responsibility for my professional actions | 0.7 | 3 | 5 | 3.98 | .68 | .02 | -.86 |
| 22. I manage health care resources efficiently | 1.0 | 2 | 5 | 3.71 | .69 | .18 | -.51 |
| 23. I make appropriate use of community resources | 11.6 | 2 | 5 | 3.38 | .73 | .15 | -.22 |
| 24. I give priority to urgent requests | 0.2 | 3 | 5 | 4.11 | .65 | -.11 | -.65 |
| 25. I handle emergency situations effectively | 5.9 | 2 | 5 | 3.93 | .73 | -.01 | -.82 |
| 26. I manage my stress effectively | 1.2 | 1 | 5 | 3.51 | .77 | .36 | -.22 |
| 27. I participate in a system of call | 7.9 | 2 | 5 | 3.90 | .74 | .08 | -1.02 |
| 28. I recognize my limitations | 1.0 | 2 | 5 | 3.78 | .68 | .26 | -.79 |
| 29. I handle requests for consultation timely | 0.7 | 1 | 5 | 3.82 | .77 | -.24 | -.13 |
| 30. Advise if referral request outside scope of practice | 1.2 | 2 | 5 | 3.79 | .73 | .23 | -.86 |
| 31. I assume appropriate responsibility for patients | 0.5 | 3 | 5 | 3.92 | .69 | .11 | -.91 |
| 32. I provide timely info to referring physicians | 0.5 | 2 | 5 | 3.80 | .75 | -.16 | -.34 |
| 33. I critically evaluate the medical literature | 0.2 | 2 | 5 | 3.91 | .75 | -.06 | -.71 |
| 34. I facilitate the learning of others | 1.5 | 2 | 5 | 3.84 | .76 | -.00 | -.71 |
| 35. I contribute to QI programs and guidelines | 10.1 | 1 | 5 | 3.59 | .89 | -.31 | -.10 |
| 36. I participate effectively as a team member | 0.7 | 2 | 5 | 3.92 | .67 | .04 | -.67 |
| 37. Professional/ethical behavior towards colleagues | 0 | 2 | 5 | 3.96 | .68 | .00 | -.71 |
| Overall Self Survey Score | | 2.89 | 5 | 3.91 | .51 | .15 | -.49 |

Note. Some items have been abbreviated. See Appendix D for complete survey

**Table 27a** Item Descriptive Statistics for Self Survey, Iteration 2

| Iteration 2 (n = 404 ) *Item* | *% Missing* | *Min* | *Max* | *Mean* | *S.D.* | *Skew* | *Kurtosis* |
|---|---|---|---|---|---|---|---|
| 1. I communicate effectively with patients | 0.0 | 3 | 5 | 4.27 | .65 | -.32 | -.72 |
| 2. I communicate effectively with patients' families | 1.0 | 2 | 5 | 4.13 | .65 | -.19 | -.44 |
| 3. I communicate effectively with other professionals | 0.0 | 3 | 5 | 4.08 | .65 | -.08 | -.61 |
| 4. I communicate treatment options to patients | 0.5 | 3 | 5 | 4.20 | .65 | -.24 | -.72 |
| 5. I perform technical procedures skillfully | 10.6 | 2 | 5 | 4.01 | .75 | -.09 | -1.0 |
| 6. I select diagnostic tests appropriately | 0.0 | 3 | 5 | 4.04 | .67 | -.05 | -.76 |
| 7. I critically assess diagnostic information | 0.2 | 3 | 5 | 4.10 | .69 | -.13 | -.87 |
| 8. I make the correct diagnosis following consultation | 0.5 | 3 | 5 | 4.04 | .65 | -.04 | -.62 |
| 9. I select appropriate treatments | 0.5 | 3 | 5 | 4.08 | .65 | -.08 | -.63 |
| 10. I maintain quality medical records | 0.2 | 2 | 5 | 3.96 | .71 | -.02 | -.82 |
| 11. I handle transfer of care appropriately | 2.5 | 3 | 5 | 3.90 | .68 | .13 | -.82 |
| 12. Clear about who is responsible for continuing care | 0.2 | 2 | 5 | 3.91 | .71 | .08 | -.88 |
| 13. I recognize psychosocial aspects of illness | 0.5 | 2 | 5 | 3.93 | .70 | .06 | -.85 |
| 14. I maintain confidentiality of patients | 1.0 | 3 | 5 | 4.16 | .71 | -.25 | -1.02 |
| 15. I co-ordinate care effectively with others | 0.5 | 2 | 5 | 3.97 | .68 | -.01 | -.70 |
| 16. I manage patients with complex problems | 1.5 | 2 | 5 | 4.11 | .70 | -.19 | -.76 |
| 17. I respect the rights of patients | 0.0 | 3 | 5 | 4.15 | .70 | -.21 | -.96 |
| 18. I show compassion for patients and their families | 0.2 | 3 | 5 | 4.13 | .71 | -.19 | -1.0 |
| 19. I collaborate with physician colleagues | 0.2 | 2 | 5 | 4.02 | .69 | -.07 | -.72 |
| 20. I am involved with professional development | 1.7 | 2 | 5 | 3.91 | .71 | .00 | -.71 |
| 21. I accept responsibility for my professional actions | 0.7 | 3 | 5 | 4.08 | .68 | -.10 | -.85 |
| 22. I manage health care resources efficiently | 1.0 | 2 | 5 | 3.84 | .68 | .11 | -.67 |
| 23. I make appropriate use of community resources | 11.6 | 2 | 5 | 3.59 | .67 | .43 | -.45 |
| 24. I give priority to urgent requests | 0.2 | 2 | 5 | 4.14 | .74 | -.30 | -.89 |
| 25. I handle emergency situations effectively | 5.9 | 2 | 5 | 3.93 | .72 | .06 | -.94 |
| 26. I manage my stress effectively | 1.2 | 1 | 5 | 3.57 | .73 | .28 | -.18 |
| 27. I participate in a system of call | 7.9 | 2 | 5 | 3.97 | .75 | -.07 | -.92 |
| 28. I recognize my limitations | 1.0 | 2 | 5 | 3.90 | .70 | .09 | -.82 |
| 29. I handle requests for consultation timely | 0.7 | 2 | 5 | 3.88 | .73 | -.08 | -.55 |
| 30. Advise if referral outside scope of practice | 1.2 | 2 | 5 | 3.93 | .70 | .06 | -.84 |
| 31. I assume appropriate responsibility for patients | 0.5 | 3 | 5 | 4.01 | .67 | -.01 | -.79 |
| 32. Provide timely information to referring physicians | 0.5 | 2 | 5 | 3.89 | .73 | .10 | -.96 |
| 33. I critically evaluate the medical literature | 0.2 | 2 | 5 | 3.94 | .74 | -.02 | -.87 |
| 34. I facilitate the learning of others | 1.5 | 2 | 5 | 3.91 | .74 | -.14 | -.52 |
| 35. I contribute to QI programs and guidelines | 10.1 | 1 | 5 | 3.69 | .84 | -.23 | -.21 |
| 36. I participate effectively as a team member | 0.7 | 3 | 5 | 3.95 | .68 | .06 | -.82 |
| 37. Exhibit professional and ethical behavior | 0 | 3 | 5 | 4.08 | .69 | -.10 | -.87 |
| Overall Self Survey Score | | 2.92 | 5 | 3.99 | .53 | -.01 | -.60 |

Note. Some items have been abbreviated. See Appendix D for complete survey.

*4.5.2 Descriptive Statistics: Attribute Scores*

Descriptive statistics for attribute scores (identified by the Exploratory Factor Analysis) are shown in Table 28. Attribute scores were lower than for the medical colleague survey. Whereas medical colleague attribute subscores were negatively skewed, self attribute cores were normally distributed.

**Table 28** Descriptive Statistics for Self Attributes Scores

| *Attribute* | *n* | *Min* | *Max* | *Mean* | *S.D.* | *Skew* | *Kurtosis* |
|---|---|---|---|---|---|---|---|
| **Iteration 1** | | | | | | | |
| Professionalism | 401 | 2.92 | 5.00 | 3.90 | .53 | .152 | -.533 |
| Clinical Competency | 401 | 3.00 | 5.00 | 4.02 | .54 | -.026 | -.672 |
| Psychosocial/Communication | 401 | 2.78 | 5.00 | 3.98 | .53 | .038 | -.465 |
| Professional Development | 400 | 2.17 | 5.00 | 3.77 | .60 | .093 | -.498 |
| **Iteration 2** | | | | | | | |
| Professionalism | 404 | 2.92 | 5.00 | 3.98 | .55 | .012 | -.625 |
| Clinical Competency | 404 | 3.00 | 5.00 | 4.07 | .55 | -.163 | -.644 |
| Psychosocial/Communication | 404 | 2.88 | 5.00 | 4.05 | .54 | -.152 | -.648 |
| Professional Development | 404 | 2.33 | 5.00 | 3.84 | .59 | .049 | -.527 |

*4.5.3 Exploratory Factor Analysis*

Principal components analysis was performed on the 37 survey items. The survey was was suitable for factor analysis, as the Kaiser measure of sampling adequacy was 0.97. Using Kaiser's Rule (Eigenvalues > 1), four factors were extracted, which explained 64.60% of the variance. The varimax-rotated pattern coefficient matrix is shown in Table 29. The majority of variables loaded on only one factor and factors were interpretable. Suggested factor labels are: Professionalism (Factor 1), Clinical Competency (Factor 2), Psychosocial/Communication (Factor 3), and Professional Development (Factor 4). These are the same four factors identified in the Medical Colleague survey.

**Table 29** Varimax-Rotated Pattern Coefficient Matrix, Self Survey

| | *Factor* | | | |
|---|---|---|---|---|
| *Item* | 1 | 2 | 3 | 4 |
| 1. I communicate effectively with patients | .201 | **.464** | **.738** | .014 |
| 2. I communicate effectively with patients' families | .208 | .329 | **.783** | .085 |
| 3. I communicate effectively with other health care professionals | **.414** | .359 | **.543** | .132 |
| 4. I communicate treatment options to patients | .269 | **.534** | **.490** | .240 |
| 5. I perform technical procedures skillfully | .355 | **.626** | .159 | .184 |
| 6. I select diagnostic tests appropriately | **.400** | **.681** | .195 | .245 |
| 7. I critically assess diagnostic information | .396 | **.684** | .236 | .298 |
| 8. I make the correct diagnosis following consultation | .375 | **.678** | .252 | .299 |
| 9. I select appropriate treatments | **.403** | **.681** | .278 | .278 |
| 10. I maintain quality medical records | **.570** | .329 | .145 | .089 |
| 11. I handle transfer of care appropriately | **.670** | .347 | .226 | .110 |
| 12. Clear about who is responsible for the continuing care of patients | **.678** | .326 | .222 | .154 |
| 13. I recognize psychosocial aspects of illness | .203 | .066 | **.720** | .277 |
| 14. I maintain confidentiality of patients and their families | **.535** | .387 | .322 | .241 |
| 15. I co-ordinate care effectively with others | **.581** | .302 | .306 | .212 |
| 16. I manage patients with complex problems | .276 | **.541** | .276 | .360 |
| 17. I respect the rights of patients | **.513** | .295 | **.486** | .277 |
| 18. I show compassion for patients and their families | **.467** | .184 | **.636** | .206 |
| 19. I collaborate with physician colleagues | **.563** | .255 | .314 | .374 |
| 20. I am involved with professional development | .319 | .224 | .138 | **.735** |
| 21. I accept responsibility for my professional actions | **.606** | .351 | .267 | .381 |
| 22. I manage health care resources efficiently | **.462** | .303 | .325 | .340 |
| 23. I make appropriate use of community resources | .360 | -.078 | **.473** | **.446** |
| 24. I give priority to urgent requests | **.588** | .344 | .243 | .177 |
| 25. I handle emergency situations effectively | **.497** | **.554** | .193 | .288 |
| 26. I manage my stress effectively | **.588** | .118 | .224 | .173 |
| 27. I participate in a system of call | **.542** | .378 | .227 | .318 |
| 28. I recognize my limitations | **.687** | .146 | .288 | .259 |
| 29. I handle requests for consultation in a timely manner | **.712** | .216 | .095 | .162 |
| 30. Advise if referral is outside the scope of my practice | **.659** | .259 | .208 | .317 |
| 31. I assume appropriate responsibility for patients | **.650** | .374 | .293 | .224 |
| 32. Provide timely information to referring physicians | **.643** | .318 | .145 | .212 |
| 33. I critically evaluate the medical literature | .282 | **.404** | .070 | **.647** |
| 34. I facilitate the learning of medical colleagues and co-workers | .212 | .306 | .220 | **.742** |
| 35. I contribute to QI programs and practice guidelines | .147 | .211 | .160 | **.797** |
| 36. I participate effectively as a member of the health care team | **.403** | .186 | **.479** | **.459** |
| 37. I exhibit professional and ethical behavior | **.640** | .267 | .361 | .251 |
| Eigenvalue | 19.58 | 1.59 | 1.50 | 1.24 |
| % Variance | 52.92 | 4.28 | 4.04 | 3.46 |

Note. Loadings > 0.4 are in bold.
Some items are abbreviated. See Appendix D for complete survey.

*4.5.4 Confirmatory Factor Analysis (CFA)*

CFA, based on the factor model derived using Exploratory Factor Analysis at Iteration 1, was performed using IBM SPSS AMOS Version 20. The main purpose of this analysis was to test the model factor structure derived at Iteration 1. A rival model, using only pure variables, was also used. For this rival model, items were only allowed to load to one factor (the one with the highest loading). For example, survey item 1 loaded to both Factor 2 (loading = 0.464) and to Factor 3 (loading = 0.738). For the initial model, both of these loadings were used. For the rival model, item 1 loaded exclusively to Factor 3. Graphics of the two models are shown in Appendix H. Model fit statistics of the initial and rival models are shown in Table 30.

**Table 30** Model Fit Statistics, Self Survey

| *Test* | *Initial Model* | *Rival Model* |
|---|---|---|
| $\chi^2$ | 2810, df = 656, p = 0.000 | 2204, df = 620, p = 0.000 |
| NFI | 0.79 | 0.83 |
| CFI | 0.83 | 0.87 |
| RMSEA | 0.09 | 0.08 |

Note. $\chi^2$ = Chi Square

NFI = Normalized Fit Index

CFI = Confirmatory Fit Index

RMSEA = Root Mean Squared Error of Approximation

### *4.5.5 Reliability*

Cronbach's alphas were calculated to determine the internal consistencies of the overall surveys, and for the attribute scales. As shown in Table 31, internal consistency was high at both iterations for the overall survey and for each attribute score.

**Table 31** Cronbach's alphas, Self Survey

| | *Iteration 1* | *Iteration 2* |
|---|---|---|
| Total Score | 0.98 | 0.98 |
| Professionalism | 0.96 | 0.97 |
| Clinical Competency | 0.93 | 0.94 |
| Psychosocial/Communication | 0.90 | 0.91 |
| Professional Development | 0.87 | 0.88 |

### *4.5.6 Change over Time*

A repeated measures multivariate analysis of variance (MANOVA) was performed to determine if there were differences in scores between Iteration 1 and 2. The hypothesis is that scores will increase from Iteration 1 to Iteration 2. The independent variable was iteration and dependent variables were the four attribute scores. Effect sizes (partial eta) were also calculated to determine magnitude of change.

Assumption testing showed that the number of cases far exceeded the number of dependent variables, and the attribute scores were normally distributed (Table 28). Using Mahalanobis distances, only 3 multivariate outliers were identified at both iterations. Assumptions of multicollinearity were violated; as shown in Table 33, the dependent variables were highly correlated with each other.

**Table 32** Correlations between Attribute Scores, Self Survey

| | Professionalism | Collaboration | Psychosocial/ Communication | Professional Development |
|---|---|---|---|---|
| Professionalism | 1 | | | |
| Clinical Competency | 0.894 | 1 | | |
| Psychosocial/Communication | 0.866 | 0.824 | 1 | |
| Professional Development | 0.772 | 0.766 | 0.751 | 1 |

Note. All correlations were significant, 2 tailed, p=0.000.

There was a significant difference on the combined dependent variables between iterations, Wilks' lambda = 0.952, F(4,396) = 4.814, p = 0.001. However, the effect size was small (partial eta squared = 0.046). All four attribute scores made unique significant contributions; Professionalism F(1, 399) = 15.50, p< 0.000; partial eta squared = 0.037, Clinical Competency F(1, 399) = 6.07, p = 0.014; partial eta squared = 0.015, Psychosocial/Communication F(1, 399) = 12.60, p = 0.000; partial eta squared = 0.031, and Professional Development F (1,399) = 6.92, p = 0.000; partial eta squared = 0.017.

Paired t-tests were also used to compare overall scores and attribute scores between iterations. The paired-t test indicated a significant difference in overall scores ($t_{400}$ = -3.73; p = 0.00), however the effect size was small (Cohen's d = 0.15). Significant increases over time were also found for the attribute scores; Professionalism $t_{400}$ = -3.82, p = 0.00, Cohen's d = 0.15 , Clinical Competency $t_{400}$ = - 2.34, p = 0.02, Cohen's d = 0.09, Psychosocial/Communication $t_{400}$ = -3.44, p = 0.001, Cohen's d = 0.13, and Professional Development $t_{399}$ = - 2.63, p = 0.009, Cohen's d = 0.12 .

*4.5.7 What Variables Predict Improvement over Iterations?*

Sequential multiple regression was performed to identify predictors of overall self survey score at Iteration 2, after controlling for overall self score at Iteration 1. The following independent variables were used: overall score at Iteration 1 for coworker, patient, and medical colleague surveys; years since graduation; location of graduation; location of practice; and sex. No major assumptions of multiple regression were violated. Sample

size was adequate and multicollinearity was absent. The independent variables were not highly correlated with each other. Additionally, tolerances were all > 0.1 and variation inflation factors were all < 10.

The multiple regression model summary is shown in Table 34. Percentage of variance of overall self score at Iteration 2 that is explained by the entire model (including overall self score at Iteration 1) is 47.5% However, after controlling for overall self score at Iteration 1, the other variables contribute less than 1% of this variance ($F(9,384) = 37.68$, $p = 0.000$). No variables made significant unique contributions.

**Table 33** Model Summary of Sequential Multiple Regression, Self Survey

| *Model* | *R* | $R^2$ | *Adjusted* $R^2$ | *S.E.* | $R^2$ *Change* | *F* | *df* | *Sig* |
|---|---|---|---|---|---|---|---|---|
| 1[a] | .683 | .466 | .465 | .384 | .466 | 334.84 | 1,384 | .000 |
| 2[b] | .689 | .475 | .462 | .385 | .008 | 37.68 | 9,384 | .000 |

a. Model 1 Predictors: Overall self score at iteration 1

b. Model 2 Predictors: All independent variables, including overall self score at iteration 1.

*4.5.8 What is the Relationship between Self and Medical Colleague Attribute Scores?*

Pearson's r correlations between self and medical colleague attribute scores at iteration 1 were calculated. Correlations were low and were not significant for professionalism ($r = 0.092$) or for clinical competency ($r = 0.096$). Correlations were statistically significant but low for psychosocial/communication score ($r = 0.157$, $p = 0.002$) and for professional development score ($r = 0.219$, $p = 0.000$).

# Chapter 5: Discussion

## 5.1 Summary of Main Findings

The current study builds on work completed almost a decade ago, which used a smaller sample (n = 103) of adult medicine specialists who participated in PAR on one occasion (Lockyer & Violato, 2004, Violato et al., 2003). The strength of the current study is the large sample size (n = 404), and the use of contemporary data. Additionally, the longitudinal nature of the current study allowed for an evaluation of change over time and for a more in depth analysis of the internal factor structure of the surveys.

Item and attribute scores were very high, and missing data rates were high for the patient survey. Exploratory factor analysis (EFA) yielded interpretable 3-4 factor solutions, which accounted for moderate to high amounts of variances. The proposed models provided for moderate fit of the data at Iteration 2. There was little or no correlation between self-assessment and medical colleagues on the same underlying attributes.

Internal consistency reliabilities of surveys were very high. Using a one-facet nested G study design, generalizability coefficients suggested moderate dependability of the assessments. Although there were statistically significant increases in scores over time, effect sizes were small, and no predictors of change were identified.

A more in depth discussion of each research question, including comparisons with previous work and hypotheses to explain the findings, follows. The chapter concludes with practical implications of our findings and suggestions for further research.

## 5.2 Research Question 1: How do coworkers, patients, and medical colleagues rate adult medicine specialists on various items and attributes?

Mean aggregate scores for all items and attributes were very high and negatively skewed on the coworker, patient, and medical colleague survey. This is in keeping with previous PAR studies of adult medicine specialists (Lockyer & Violato, 2004, Violato et al., 2003)

and of non-PAR MSF tools for practising physicians (Archer et al., 2010, Lipner et al., 2002, Wright et al., 2012). The high scores may partially be due to rater self-selection, in that participants may select raters who they suspect will give them positive assessments. For example, Archer & McAvoy (2011) found that 50% of assigned peer raters gave scores of "less than satisfactory" to physicians previously identified as performing poorly, but this dropped to 19% when raters were self-selected.

Alternative explanations for the high scores seen in the current study include the halo effect, perceived negative consequences of giving low scores, or fear that the process was not confidential (Williams, Klamen & McGaghie, 2009). The halo effect has been called "the most pervasive error in performance appraisal" and occurs when individual items of performance are influenced by the rater's overall impression of a person (Nathan & Lord, 1983, Streiner & Norman, 2008). Unfortunately, the halo effect persists despite rater training and behavioral anchors on scales (Nathan & Lord, 1983), likely because it is a fairly fixed cognitive process (Govaerts, van der Vleuten, Schuwirth & Muijtjens, 2007).

The above-mentioned potential explanations for the high scores are all examples of construct-irrelevant variance, which is defined as "the degree to which test scores are affected by processes that are extraneous to its intended construct" (American Educational Research Association, 1999). Construct-irrelevant variance threatens validity and if possible should be identified and reduced (Downing & Haladyna, 2004). It is unknown if rater training prior to participating in the PAR assessment process would reduce this variance. The available evidence on rater training suggests that it is not cost effective (Williams et al., 2012). Perhaps a more appealing option is to adjust ratings by "controlling for" leniency (Downing & Haladyna, 2004).

Many items on the patient survey had high missing data rates. For example, in Iteration 1, 38% of items had missing rates > 15%, 25% of items had missing rates > 25%, and 8% had missing rates > 40%. These high rates are similar to the findings of Violato et al.

(2003), who found 14/40 items had "unable to assess" rates of > 20%. It is possible that this represents a response bias. Mazor, Clauser, Field, Yood, and Gurwitz (2002) found a positive correlation between mean patient satisfaction rating and response rate in a patient satisfaction survey of primary care physicians. Simulation studies suggested that this response bias led to an overestimation of patient satisfaction overall, and the effect was greatest for physicians with the lowest scores. Similarly, in evaluating response biases among patients in a National Board of Medical Examiners MSF tool used with trainees, Mazor, Clauser, Holtman and Margolis (2007) found that the number of questions answered was higher for trainees with the highest overall scores and lowest for trainees with the lowest overall scores. In other words, missing responses were not random but rather were systematically related to the performance of the trainee.

Thus patient raters may be more likely to answer questions when they view the physician positively and more likely to leave questions blank or select "unable to assess" when they view the physician negatively. If true, it may contribute both to the high scores and the high rates of missing data in the patient survey in the current study. An alternative explanation is that patients did not directly observe the behavior of interest. The current adult medicine patient survey was originally modified from the family physician assessment, and the content validity has not explicitly been explored using adult medicine specialists. Some items may not be widely applicable to adult medicine specialists. This possibility is concerning, as too few observations of the behaviour of interest leads to "construct underrepresentation" which is a serious threat to content validity in ratings of clinical performance (Downing & Haladyna, 2004).

### 5.3 Research Question 2: What underlying attributes does each survey actually measure, and are these stable over time?

Using EFA with varimax-rotation and Kaiser's rule for factor extraction, the following 3-4 factor solutions were proposed based on Iteration 1 data:

- Coworker – Professionalism, Collaborator, Psychosocial/Communication

- Medical Colleague – Professionalism, Collaborator, Psychosocial/Communication, Clinical Competence
- Patient – Psychosocial/Humanistic, Communication, Clinic Staff, Office Staff
- Self – Professionalism, Psychosocial/Communication, Collaborator, Clinical Competence

These solutions explained a moderate proportion of the variance (64.6 to 72.3%). There is debate about what is an acceptable proportion of variance accounted for in factor analysis (Henson & Roberts, 2006). Some authors suggest that instruments with clear internal structure should explain 75% or more variance. The proportion of variance explained, however, will decrease when the number of items to be analyzed increases (Henson & Roberts, 2006). The PAR surveys are relatively long questionnaires, and this may explain why only a moderate proportion of the variance was explained.

The factor solutions were similar to those reported previously for PAR adult medical specialists (Lockyer & Violato, 2004; Violato et al., 2003). Slight differences should be highlighted, however. With the medical colleague and self surveys, we renamed factor 1 to be "professionalism" (rather than "psychosocial patient management"), as we felt it better represented the associated items. Additionally, whereas the PAR technical report described a 5-factor solution (Violato et al., 2003), we obtained a 4-factor solution for the patient survey. It could be argued that the fifth factor should not have been included in the former study, as the Eigenvalue was 0.87.

These three to four factor solutions in PAR surveys are inconsistent with non-PAR MSF tools, which consistently yield two factor solutions. For example, Ramsey et al. (1993), Wright et al. (2012), Archer, Norcini, Southgate, Heard and Davis (2008) and Archer et al. (2010) all reported two factor solutions: typically one factor measuring clinical competencies and one measuring psychosocial attributes. Overall, these non-PAR studies reported higher percentage variance accounted for and higher pattern coefficients compared to the current study.

Reasons for the discrepancy in the number of underlying constructs between PAR and non-PAR surveys are not entirely clear. The PAR surveys are longer than most MSF tools to assess practicing physicians, which may allow for a more sophisticated internal structure. Alternatively, it is possible that the differences are due to the methodological decisions made when performing the EFA. Unfortunately, most of the above mentioned studies did not consistently report EFA methodology in detail, making direct comparison difficult. One of the main challenges of EFA is that it largely depends on judgment for interpretation. The current study used Kaiser's rule to determine factor retention, as it is frequently used in medical education research and has consistently yielded interpretable and meaningful results with PAR data. However, Kaiser's rule is considered by some researchers to be the least accurate of factor extraction methods, based on evidence that it severely overestimates the number of factors to retain compared to other factor retention methods (Wetzel, 2012), and many experts advocate for the use of multiple criteria to select the number of factors to extract (Henson & Roberts, 2006).

In the current study, the first factor extracted for each survey explained the majority of variance in the solution, with the remaining factors contributing very little. Moreover, the factors were highly correlated with each other (Pearson's $r > 0.8$). Both of these observations suggest against three to four factor solutions. If the surveys were truly measuring unique attributes, we would expect correlations between factors of the same survey to be lower (Archer & McAvoy, 2011). Similarly, the current study also used orthogonal rotation. Some experts, however, advocate for oblique rotation, particularly if there is an expectation of correlation between the constructs (as is the case here). It is possible that using a different extraction method and different rotation method would lead to a different factor solution in the present study.

The proposed models (factor structures) developed from data at Iteration 1 were analyzed at Iteration 2 using CFA. None of the model fit statistics for the four surveys met conventional cutoff criteria for good model fit. The $\chi^2$ significance test did not support model fit, but this was expected due to our large sample sizes (Bentler & Bonnett, 1980).

The CFIs and NFIs in the current study ranged from 0.81 to 0.89, and the RMSEAs ranged from 0.08 to 0.11. A priori, we considered values of > 0.95 for CFI and NFI and < 0.06 for RMSEA as the cutoff values to accept the models as having good fit (Hu & Bentler, 1999). Despite not meeting these pre-defined cutoff criteria, the proposed models fit the data much better than the independence model (which assumes no correlation between variables). Additionally, all but one RMSEA value was < 0.1 (values > 0.1 are generally considered to indicate poor fit). Given the complexity of the current models, and the fact that no additional model specifications were done, the model fit indexes in the current study to provide moderate support for the current internal structure of the surveys.

The purpose of EFA and CFA in the current study was to determine if items *intended* to measure a given physician attribute are *actually* measuring that attribute. Thus the results are central to the validity of the PAR assessment. Our results provide moderate support for the current internal structure of the surveys, however further analyses and instrument development are still needed to further clarify what attributes each survey is actually measuring.

### 5.4 Research Question 3: Are the Current Surveys Reliable?

As expected based on previous work (Lockyer & Violato, 2004, Violato et al. 2003,), internal consistency reliability of the four surveys was high, with Cronbach's alphas ranging between 0.87 to 0.98 for all attributes. This is excellent, as Cronbach's alphas > 0.9 are acceptable for the highest of stakes examinations (Downing, 2004).

The generalizability coefficients were highest for the coworker survey (overall 0.79 and 0.78 for iteration 1 and 2, respectively), and lower for the medical colleague (overall 0.71 and 0.7) and the patient survey (0.68 and 0.70). Our findings are lower than 0.82 previously reported for the medical colleague survey (Lockyer & Violato, 2004), and generalizability coefficients for the coworker and patient surveys have never previously been reported. Traditionally, a cutoff of greater than 0.8 is used in the majority of

health profession studies (Narayan, Greco & Campbell, 2010). Only the coworker survey was close to meeting this cutoff. However, our results are in keeping with other studies of MSF in practicing physicians (Campbell et al., 2008, Lipner et al., 2002, Ramsey et al., 1993).

Recently, the appropriateness of using classical generalizability theory in multisource feedback of physician performance has been questioned (Narayanan, Greco & Campbell, 2010). Classical generalizability theory assumes the same number of raters for all participants, and identical raters for all participants on both occasions. These assumptions are not met in the PAR assessment process. Although there *proposed* number of raters for each survey, the *actual* rater number differs. Additionally, raters are likely unique to each participant and to each iteration. Modified formulas for generalizability theory and D studies have been proposed, which account for these assumption violations (Narayanan et al., 2010). A comparison of generalizability estimations using these two methods is an area for future research.

### 5.5 Research Question 4: Do scores improve over time, and if yes, can predictors of those changes be identified?

For all surveys there was a statistically significant increase in scores between iterations. However, the effect sizes for the self, coworker, and patient surveys were small. After controlling for initial iteration 1 scores on the survey of interest, no unique predictors (such as demographic factors or scores on other surveys at iteration1) were identified that contributed in a meaningful way to scores at iteration 2. A PAR longitudinal study of family physicians also found a statistically significant increase between iterations for the medical colleague survey (moderate effect size) and coworker survey (small effect size), but not for the patient survey (Violato, Lockyer & Fidler, 2008).

An underlying assumption of MSF is that it has a "catalytic effect" and stimulates further learning (Norcini et al., 2011). Evidence of change over time would provide evidence in support of consequential validity (American Educational Research Association, 1999).

The small effect sizes seen in the current study may be due to a lack of educational impact of the PAR assessment. Overeem et al. (2010) argues that "simple feedback" (in this case, a feedback report) does not work to improve performance, and called for a portfolio that stimulated reflection and a trained facilitator to deliver the MSF. Alternatively, it is possible that physicians do not value feedback from patients and coworkers (adult medicine specialists' perceptions of PAR have never been explored). Finally, Lockyer et al., (2011) found that physicians are more open to feedback that originated from activities of their choosing (such as non-formal meetings/discussions with peers), and more resistant to feedback from mandatory participation (such as the PAR program).

However, there are other equally plausible explanations for the lack of meaningful increase in scores between iterations. First, the "ceiling effect" of scores may prevent statistical detection of change (Streiner & Norman, 2008). Moreover, although participants' scores are compared to others (participants are given their score "rank" compared to other adult medicine specialists on the same attributes), the influence of these comparisons on feedback interpretation is unknown. Participants are also given absolute item and attribute scores as part of their feedback, and these values are high. If the absolute scores influence feedback interpretation more than rankings, physicians may not feel compelled to change, as they interpret that they are doing well. Furthermore, the current study focused on global changes in overall and attribute scores, making it insensitive to major changes in one or two specific areas. Significant changes in one or two domains (items) may be a realistic and acceptable performance goal between one assessment to another, but these would not necessarily be detected by global or attribute ratings (Smither, London, & Reilly, 2005).

The degree of improvement over time was highest for the medical colleague survey (effect sizes were small to moderate). This was also found in a PAR longitudinal study of family physicians (Violato et al., 2008). This increase may reflect true performance change. If it does, it suggests the adult medicine specialists value feedback from peers

more than from coworkers or patients, as found by others (Overeem et al., 2012). An alternative explanation is that "experienced" participants are more careful at the second iteration to select assessors that are likely to give positive reviews.

Self-assessment scores were fairly stable over five years. Although there was a statistically significant increase in scores between iterations, the effect sizes were small and thus unlikely to be particularly meaningful. Almost half of the overall self-assessment score variance at Iteration 2 could be explained by self-assessment scores at Iteration 1. Other factors, such as demographics or scores of other surveys at Iteration 1, did not influence self- assessment scores at Iteration 2. This is in keeping with the work of Lockyer et al., (2007), who also found that socio-demographic factors and Iteration 1 attribute scores from other surveys did not predict self-assessment scores at Iteration 2.

### 5.6 Research Question 5: What is the relationship between self-assessment attribute scores and corresponding medical colleague attribute scores?

With the exception of one item, the self-assessment and medical colleague surveys are identical. Therefore it is not surprising that the same four underlying attributes were identified: Professionalism, Clinical Competence, Psychosocial/Communication and Professional Development. The loading matrices were similar (although not identical), and the proportion of variance accounted for by each attribute was similar. However, the correlations between medical colleagues and self-assessment scores for the same attribute were low (ranging from 0.092 for Professionalism to 0.22 for Professional Development). This lack of correlation is consistent with previous PAR publications (Lockyer et al., 2007). Physicians are typically more accurate in assessing peers than they are in assessing their own abilities, using a third external assessment as the "gold standard" (Colthart et al., 2008). It is plausible that the lack of/low correlation found in the current study is due to inaccurate self-assessment. However, because the true "gold standard" assessment is unknown, this conclusion cannot definitively be made.

Compared to the other three surveys, the self-assessment data was unique in several ways. First, data was normally distributed, suggesting that a "ceiling effect" does not exist for self-assessments. Additionally, overall and attribute scores were lower, also suggesting against a "ceiling effect" for self-assessment. Finally, missing data rates were much lower, suggesting that the survey is acceptable and feasible for participants.

### 5.7 Limitations

The conclusions made in the current chapter must be understood in the context of the limitations of the current study. As mentioned previously, the self-selection of raters is a potential limitation, particularly if it is contributing to score inflation. Additionally, we know very little about the patient raters. For example, demographic information is limited and the severity of underlying illness and duration of the physician/patient relationship is unknown. Moreover, with the exception of subspecialty, relatively little is known about the clinical practice of the physician. Specifically, it is unknown if the physician works predominantly in an inpatient or outpatient setting, and it is unknown from which of these two settings raters are being selected.

The study was longitudinal, but due to design of the PAR process we were limited to assessment 5 years apart. It is possible that more improvement would occur if the assessment cycle was shortened. Finally, we are unaware of how adult medicine specialists interpret or value their feedback report.

### 5.8 Practical Implications for PAR

The findings presented here have immediate practical implications for assessing adult medicine specialists using MSF. The underlying attributes identified in the present study only partially reflect those stated on the PAR website and in the feedback report given to participants. For example, the PAR website states that one of the three attributes measured by the Coworker Surveys is called "Patient Interaction". The following is the PAR description of this attribute (with the corresponding survey item in parenthesis):

"The medical specialist gives patient reasonable access (item 14) and communicates effectively with them and their families (items 1 and 17) in a non-judgmental manner than conveys respect and compassion (items 9). The medical specialist responds appropriately in an emergency situation (items 19), maintains confidentiality (item 15) and accepts responsibility for professional actions (item 18)".

The present study, as well as the findings of Violato et al (2003), suggest that these specific coworker items reflect two, rather than one, underlying attributes: items 1, 9, and 17 reflect "Psychosocial/Communication", and items 15, 18, and 19 reflect "Professionalism". Additionally, the PAR website states that the patient survey measures six attributes, but our results suggest only four underlying attributes are measured. The PAR program should consider revisiting the listed attributes on the feedback report and website to more accurately reflect the underlying attributes identified by factor analysis.

### 5.9 Suggestions for Future Research

Findings of the current study pose the following questions that warrant further exploration:

*1. Does rater self-selection threaten validity in the PAR assessment?*

The decision to allow self-selection of raters in the PAR assessment was based on the work of Ramsey et al. (1993), who found no difference in physician performance ratings from self-selected versus assigned raters. Recent evidence suggests, however, that assigned medical colleague raters provided significantly lower scores than self-selected raters (Archer & McAvoy, 2011). It is possible that this self-selection contributes to the high scores seen in the current study, and even to the improvement of scores over time: physicians may get "better" at selecting their raters the second time around. If such a bias is present, self-selection of raters could pose a serious threat to score validity. A study comparing self-selected versus assigned raters is needed.

2. *Can we eliminate the high scores?*

Rater training is often suggested to reduce high scores, although previous research suggests that it is not cost effective nor does it eliminate the halo effect. As mentioned in the previous paragraph, assigning raters may decrease score inflation. Shorter rating scales may decrease scores in MSF, but would require complete revision of the PAR surveys (Hassell et al., 2012). A practical option worthy of further investigation is statistical transformation of scores to in order to "control for" leniency. Using PAR data from pathologists and lab medicine specialists, Violato and Lockyer (2013) recently showed that PAR data can be normalized using the natural logarithm followed by T-score transformation. The resulting score spread allows for differentiation between levels of performance, which the authors suggest will contribute to enhanced understanding and acceptance of the feedback report.

3. *How can the patient survey be improved*?

Of the four surveys, the patient survey has the most validity-related concerns. Specifically, scores are very high and negative skewed, missing data rates are concerning, and 24 raters ensure only moderate assessment dependability. A critical analysis of all of the items, by a representative group of adult medicine specialists and their patients, is needed.

With regards to the patient survey, patient characteristics and contextual factors may need to be considered as there evidence that they influence ratings (Duberstein, Meldrum, Fiscell, Shields & Epstein, 2007, Govaerts et al., 2007, Norcini, 2005, Campbell et al., 2011). Lipner et al. (2002) found that healthier patients and those who knew the physician longer, tended to rate their internal medicine specialist higher on MSF. Many internists work in both inpatient and outpatient settings, and currently the environment from which the patient raters are recruited in PAR is unknown. Sick patients in an inpatient setting (who have often just met the specialists) observe different behaviors and likely hold different values than patients in an outpatient clinic setting (who are generally more healthy and may have a longer relationship with the physician). Collecting more

characteristics about the patient raters in the PAR assessment may provide insight into the ratings.

4. *What is the educational impact of PAR on adult medicine specialists?*

This question should be a priority for future research, as physician improvement is the primary goal of the PAR assessment process. In the current study, scores were fairly stable over time, however we do not know if this is a shortcoming in the ability to statistically detect a change, or because PAR lacks an educational impact. Qualitative studies focusing on the interpretation and perceived value of the feedback report in adult medicine specialists would be informative. Manipulating factors previously found to improve the catalytic effect of feedback – for example timing between assessments and delivery methods - is also warranted.

5. *How do the modified formulas for generalizability theory compare to classical generalizability in the current sample?*

As mentioned above, these new formulae account for the assumption violations of classical generalizability theory (Narayanan et al., 2010), but have never been used in previous PAR publications.

6. *How can the factor structure be improved?*

A detailed analysis of CFA model specification/respecification is needed to understand the limitations of our current proposed factor structures. Next, a critical look at the decisions used in the EFA is warranted. For example, it is possible that the use of statistic factor scores (rather than crude factor scores), oblique rotation (rather than orthogonal rotation), and using multiple criteria to determine the number of factors to extract will improve model fit. Finally, a revision of survey content may be warranted.

7. *What is the relationship between PAR assessments and external observations of performance?*

Evidence of current or predictive validity of MSF would certainly strengthen the overall validity of the entire assessment process. To date, this area has been largely unexplored, in part due to the paucity of regular, standardized assessment of physicians in practice. Only two published study has explored the relationship of a MSF program for practicing physicians to external measures of performance (Archer & McAvoy, 2011, Wright et al., 2012). Postgraduate trainees may be the best population to further establish the concurrent and predictive validity of MSF, as there are many other assessments/milestone achievements with which to draw comparisons. Unfortunately, whether or not validity in the trainee population will successfully generalize to practicing physicians is unknown.

## 5.10 Conclusion

The present study builds on the existing foundation of validity-related evidence to support the PAR assessment of adult medicine specialists. It also informs the current applicability of the surveys, which have been unchanged in the last decade. The strength of this study is the large sample size (n = 404), and the use of current data. Additionally, the longitudinal nature allowed for an evaluation of change over time and for a comprehensive analysis of the internal factor structure of the surveys.

This study demonstrated that scores on all surveys were high and negatively skewed. The patient survey had a high rate of missing data. All surveys had high internal consistency reliability and moderate generalizability. The three to four factor solutions proposed using exploratory factor analysis at Iteration 1 provided for moderate fit of the data at Iteration 2. There was a significant increase in scores over time for all four surveys, with small to moderate effect sizes. There was little or no correlation between self-assessment and medical colleagues on the same underlying attributes.

These findings support adequate generalizability of the PAR assessment for adult medicine specialists. However, the score inflation and lack of model fit using

confirmatory factor analyses pose validity-related concerns. Therefore, we support the view of Wright et al. (2012), that PAR should be viewed as a formative, rather than summative, assessment method. All surveys had statistically significant increases in scores over time, although the effect sizes were generally small. Because the primary goal of PAR is to provide formative feedback to physicians, factors influencing the lack of meaningful change over time warrant further investigation.

## References

American Educational Research Association (Eds.). (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association, American Psychological Association, & National Council on Measurement in Education.

Archer J.C., & McAvoy P. (2011). Factors that might undermine the validity of patient and multi-source feedback. *Medical Education*, 45, 886-893.

Archer J., McGraw M., & Davies H. (2010). Republished paper: Assuring validity of multisource feedback in a national programme. *Postgraduate Medical Journal*, 86, 526-531.

Archer J., Norcini J., Southgate L., Heard S., & Davies H. (2008). Mini-PAT (peer assessment tool): A valid component of a national assessment programme in the UK? *Advances in Health Sciences Education*, 13(2), 181-192.

Bentler P.M. & Bonett D.G. (1980). Significance Tests and Goodness of Fit in the Analysis of Covariance Structures. *Psychological Bulletin*, 88(3), 588-606.

Brinkman W.B., Geraghty S.R., Lanphear B.P., Khoury J.C., Gonzalez del Rey J.A., Dewitt T.G., & Britto M.T. (2007). Effect of multisource feedback on resident communication skills and professionalism: a randomized controlled trial. *Archives of Pediatric and Adolescent Medicine*, 161(1), 44-49.

Butterfield P.S., & Mazzaferri E.L. (1991). A new rating form for use by nurses in assessing residents' humanistic behavior. *Journal of General Internal Medicine*, 6(2), 155-161.

Campbell J.L., Richards S.H., Dickens A., Greco M., Narayanan A., & Brearley S. (2008). Assessing the professional performance of UK doctors: an evaluation of the utility of the General Medical Council patient and colleague surveys. *Quality and Safety in Health Care*, 17(3), 187-193.

Campbell J.L., Roberts M., Wright C., Hill J., Greco M., Taylor M., & Richards S. (2011). Factors associated with variability in the assessment of UK doctors' professionalism: Analysis of survey results. *British Medical Journal*, 343, d6212 doi: 10.1136/bmj.d6212.

Colthart I., Bagnall G., Evans A., Allbutt H., Haig A., Illing J., & McKinstry B. (2008). The effectiveness of self-assessment on the identification of learner needs, learner activity, and impact on clinical practice: BEME Guide no. 10. *Medical Teacher*, 30(2), 124-145.

Cook, D.A., & Beckman, T.J. (2006). Current concepts in validity and reliability for psychometric instruments: Theory and application. *The American Journal of Medicine*, 119, 166, e7-166.e16.

Cope D.W., Linn L.S., Leake B.D., & Barrett P.A. (1986). Modification of residents' behavior by preceptor feedback of patient satisfaction. *Journal of General Internal Medicine*, 1(6), 394-398.

Davis D.A., Mazmanian P.E., Fordis M., Van Harrison R., Thorpe K.E., & Perrier L. (2006). Accuracy of physician self-assessment compared with observed measures of competence: a systematic review. *Journal of the American Medical Association*, 296(9), 1094-1102.

Downing S.M. (2003). Validity: On meaningful interpretation of assessment data. *Medical Education*, 37(9), 830-837.

Downing S.M. (2004). Reliability: On the reproducibility of assessment data. *Medical Education*, 38(9), 1006-1012.

Downing S.M., & Haladyna T.M. (2004). Validity threats: Overcoming interference with proposed interpretations of assessment data. *Medical Education*, 38(3), 327-333.

Duberstein P., Meldrum S., Fiscella K., Shields C.G., & Epstein, R. M. (2007). Influences on patients' rating of physicians: Physicians demographics and personality. *Patient Education and Counseling*, 65, 270-274.

Eva, K.W., & Regehr, G. (2005). Self-Assessment in the Health Professions: A Reformulation and Research Agenda. *Academic Medicine*, 80(10), S46-54.

Evans R.G., Edwards A., Evans S., Elwyn B., & Elwyn G. (2007). Assessing the practising physician using patient surveys: A systematic review of instruments and feedback methods. *Family Practioner*, 24(2), 117-127.

Fan X., Thompson B., & Wang L. (1999). Effects of sample size, estimation methods,a nd model specification on structural equation modeling fit indexes. *Structural Equation Modeling*, 6(1), 56-83.

Fidler H., Lockyer J.M., Toews J., & Violato C. (1999). Changing physicians' practices: the effect of individual feedback. *Academic Medicine*, 74(6), 702-714.

Govaerts M.J., van der Vleuten C.P., Schuwirth L.W., & Muijtjens A.M. (2007). Broadening perspectives on clinical performance assessment: Rethinking the nature of in-training assessment. *Advances in Health Sciences Education*, 12(2), 239-260.

Hall W., Violato C., Lewkonia R., Lockyer J., Fidler H., Toews J., Jennett P., et al. (1999). Assessment of physician performance in Alberta: The physician achievement review. *Canadian Medical Association Journal*, 161(1), 52-57.

Hassell A., Bullock A., Whitehouse A., Wood L., Jones P., & Wall D. (2012). Effect of rating scales on scores given to junior doctors in multi-source feedback. *Postgraduate Medical Journal*, 88(1035), 10-14.

Henson R.K., & Roberts J.K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement,* 66(3), 393-416.

Hu L.T. & Bentler P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1-55.

Linn L.S., Oye R.K., Cope D.W., & DiMatteo M.R. (1986). Use of nonphysician staff to evaluate humanistic behavior of internal medicine residents and faculty members. *Journal of Medical Education,* 61(11), 918-920.

Lipner R.S., Blank L.L., Leas B.F., & Fortna G.S. (2002). The value of patient and peer ratings in recertification. *Academic Medicine*, 77(10), S64-66.

Lockyer J., Armson H., Chesluk B., Dornan T., Holmboe E., Loney E., Mann K., et al. (2011). Feedback data sources that inform physician self-assessment. *Medical Teacher*, 33(2), 113-120.

Lockyer J., & Violato C. (2004). An examination of the appropriateness of using a common peer assessment survey to assess physician skills across specialties. *Academic Medicine*, 79(10), S5-8.

Lockyer J.,Violato C., & Fidler H. (2003). Likelihood of change: a study assessing surgeon use of multisource feedback data. *Teaching and Learning in Medicine,* 15(3), 168-174.

Lockyer J.M., Violato C., & Fidler H. (2006). The assessment of emergency physicians by a regulatory authority. *Academic Emergency Medicine,* 13(12), 1296-1303.

Lockyer J.M., Violato C., & Fidler H. (2006). A multi source feedback program for anesthesiologists. *Canadian Journal of Anaesthesiology*, 53(1), 33-39.

Lockyer J.M., Violato C., & Fidler H.M. (2007). What multisource feedback factors influence physician self assessments? A five-year longitudinal study. *Academic Medicine,* 82(10), S77-80.

Lockyer J.M., Violato C., & Fidler H.M. (2008). Assessment of radiology physicians by a regulatory authority. *Radiology*, 247(3), 771-778.

Lockyer J.M., Violato C., Fidler H., & Alakija P. (2009). The assessment of pathologists/laboratory medicine physicians through a multisource feedback tool. *Archives of Pathology and Laboratory Medicine*, 133(8), 1301-1308.

Mazor K.M., Clauser B.E., Field T., Yood R.A., & Gurwitz J.H. (2012). A demonstration of the impact of response bias on the results of patient satisfaction surveys. *Health Services Research*, 37(5), 1403-1417.

Mazor K., Clauser B.E., Holtman M., & Margolis M.J. (2007). Evaluation of missing data in an assessment of professional behaviors. *Academic Medicine*, 82(10), S44-47.

Mitchell C., Bhat S., Herbert A., & Baker P. (2011). Workplace-based assessment of junior doctors: Do scores predict training difficulties? *Medical Education*, 45, 1190-1198.

Murphy D.J., Bruce D.A., Mercer S.W., & Eva K.W. (2009). The reliability of workplace-based assessment in postgraduate medical education and training: A national evaluation in general practice in the United Kingdom. *Advances in Health Science Education*, 14(2), 219-232.

Nathan B.R., & Lord R.G. (1983). Cognitive categorization and dimensional schemata: A process approach to the study of halo in performance ratings. *Journal of Applied Psychology*, 68(1), 102-114.

Narayanan A., Greco M., & Campbell J.L. (2010). Generalizability in unbalanced, uncrossed and fully nested studies. *Medical Education*, 44(4), 367-378.

Norcini, J.J. (2005). Current perspectives in assessment: the assessment of performance at work. *Medical Education*, 39, 880-889.

Norcini J., Anderson B., Bollela V., Burch V., Costa M.J., Duvivier R, Galbraith, R., et al. (2011). Criteria for good assessment: Consensus statement and recommendations from the Ottawa 2010 Conference. *Medical Teacher*, 33(3), 206-214.

Norcini J., & Burch V. (2007). Workplace-based assessment as an educational tool: AMEE Guide No. 31. *Medical Teacher*, 29(9), 855-871.

Overeem K., Lombarts M.J., Arah O.A., Klazinga N.S., Grol R.P., & Wollersheim H.C. (2010). Three methods of multi-source feedback compared: A plea for narrative comments and coworkers' perspectives. *Medical Teacher*, 32, 141-147.

Overeem K., Wollersheim H.C., Arah O.A., Crujsberg J.K., Grol R., & Lombarts K. (2012). Factors predicting doctors' reporting of performance change in response to multisource feedback. *BMC Medical Education*. 2012, 12(52), 1-7.

Overeem K., Wollersheim H.C., Arah O.A., Crujsberg J.K., Grol R. & Lombarts K. (2012). Evaluation of physicians' professional performance: An iterative development and validation study of multisource feedback surveys. *BMC Health Services Research*. 12(80), 80-90.

PAR Physician Achievement Review. (n.d.). Retrieved March 9, 2013 from http://www.par-program.org/.

Ramsey P.G., Wenrich M.D., Carline J.D., Inui T.S., Larson E.B., & LoGerfo J.P. (1993). Use of peer ratings to evaluate physician performance. *Journal of the American Medical Association*, 269(13), 1655-1660.

Sargeant J.M., Mann K.V., Ferrier S.N., Langille D.B., Muirhead P.D., Hayes V.M., Sinclair D.E. (2003). Responses of rural family physicians and their colleague and coworker raters to a multi-source feedback process: a pilot study. *Academic Medicine*, 78(10), S42-44.

Smither J.W., London M., & Reilly R.R. (2005). Does performance improve following multisource feedback? A theoretical model, meta analysis, and review of empirical findings. *Personnel Psychology*, 58(1), 33-66.

Tabachnick B.G. & Fidell L.S. (2007). *Using Multivariate Statistics*. (5th ed.). Pearson Education, Inc.

Thompson, B. (2004). *Exploratory and Confirmatory Factor Analysis: Understanding Concepts and Applications.* Washington, DC: American Psychological Association.

Violato C., & Hecker K.G. (2007). How to use structural equation modeling in medical education research: A brief guide. *Teaching and Learning in Medicine*, 19(4), 362-371.

Violato C., & Lockyer J. (2006). Self and peer assessment of pediatricians, psychiatrists and medicine specialists: implications for self-directed learning. *Advances in Health Sciences Education,* 11(3), 235-244.

Violato C., & Lockyer J. (2013). Individual reporting of multi-source feedback data: Normalization and transformation to T-scores. Manuscript submitted for publication.

Violato C., Lockyer J.M., & Fidler H. (2006). Assessment of pediatricians by a regulatory authority. *Pediatrics,* 117(3), 796-802.

Violato C., Lockyer J.M., & Fidler H. (2008). Assessment of psychiatrists in practice through multisource feedback. *Canadian Journal of Psychiatry*, 53(8), 525-533.

Violato C., Lockyer J.M., & Fidler H. (2008). Changes in performance: a 5-year longitudinal study of participants in a multi-source feedback programme. *Medical Education*, 42(10), 1007-1013.

Violato C., Lockyer J., Toews J., & Fidler H. (2003). The physician achievement review program: A pilot study for Alberta specialist physicians (Technical Report*). Copy available on request.*

Wetzel A. (2012). Factor analysis methods and validity evidence: A review of instrument development across the medical education continuum. *Academic Medicine*, 87(8), 1-10.

Williams R.G., Klamen D.A., & McGaghie W.C. (2003). Cognitive, social and environmental sources of bias in clinical performance ratings. *Teaching and Learning in Medicine: An International Journal*, 15(4), 270-292.

Wright C., Richards S.H., Hill J.J., Roberts M.J., Norman G.R., Greco M., Taylor M.R., & Campbell J.L. (2012). Multisource feedback in evaluating the performance of doctors: The example of the UK general medical council patient and colleague surveys. *Academic Medicine,* 87(12), 1668-1678.

# Appendix A: Coworker Survey

3921201211

College of Physicians and Surgeons
Province of Alberta

## Co-Worker Questionnaire

PAR
Physician Achievement Review

Assessed Physician's Name: Dr.

Your Name:

**Marking Instructions**

Please indicate your answer by filling in the bubble *like this,* ● not like ⊗or ✓ . Thank you!

This form is used by a variety of physicians' co-workers (e.g. nurses, pharmacists, psychologists), therefore, not all of the following items may be relevant to you. If any of the items are **NOT** relevant to you, mark these "**Unable to Assess.**"

**How well do you know this physician (Mark one)?**

○ Not at All ○ Not Well ○ Somewhat ○ Well ○ Very Well

**Interpretation of the Rating Scale**
Please rate this physician on each of the performance statements listed according to the following scale.

**Compared to other medical specialists I know, this one is:**

| | Among the Worst 1 | Bottom Half 2 | Average 3 | Top Half 4 | Among the Best 5 | Unable to Assess UA |
|---|---|---|---|---|---|---|
| 1. Communicates effectively with patients | ○ | ○ | ○ | ○ | ○ | ○ |
| 2. Is able to verbally communicate effectively with other health care professionals | ○ | ○ | ○ | ○ | ○ | ○ |
| 3. Is able to effectively communicate in writing with other health care professionals | ○ | ○ | ○ | ○ | ○ | ○ |
| 4. Writes legibly | ○ | ○ | ○ | ○ | ○ | ○ |
| 5. Is courteous to co-workers | ○ | ○ | ○ | ○ | ○ | ○ |
| 6. Demonstrates appropriate concern for co-worker safety | ○ | ○ | ○ | ○ | ○ | ○ |
| 7. Respects the professional knowledge and skills of co-workers | ○ | ○ | ○ | ○ | ○ | ○ |
| 8. Collaborates well with co-workers | ○ | ○ | ○ | ○ | ○ | ○ |
| 9. Shows compassion to patients and their families | ○ | ○ | ○ | ○ | ○ | ○ |
| 10. Is able to separate personal values from the management of patients | ○ | ○ | ○ | ○ | ○ | ○ |
| 11. Is courteous to patients and their families | ○ | ○ | ○ | ○ | ○ | ○ |
| 12. Respects the rights of patients to make informed decisions | ○ | ○ | ○ | ○ | ○ | ○ |
| 13. Accepts responsibility for patient care | ○ | ○ | ○ | ○ | ○ | ○ |
| 14. Is reasonably accessible to patients | ○ | ○ | ○ | ○ | ○ | ○ |
| 15. Maintains confidentiality of patients | ○ | ○ | ○ | ○ | ○ | ○ |
| 16. Is accessible for appropriate communication about patients | ○ | ○ | ○ | ○ | ○ | ○ |
| 17. Communicates effectively with families | ○ | ○ | ○ | ○ | ○ | ○ |
| 18. Accepts responsibility for professional actions | ○ | ○ | ○ | ○ | ○ | ○ |
| 19. Responds appropriately in emergency situations | ○ | ○ | ○ | ○ | ○ | ○ |
| 20. Participates effectively as a member of the health care team | ○ | ○ | ○ | ○ | ○ | ○ |
| 21. Facilitates the learning of co-workers | ○ | ○ | ○ | ○ | ○ | ○ |
| 22. This doctor presents him/herself in a professional manner | ○ | ○ | ○ | ○ | ○ | ○ |

# Appendix B: Medical Colleague Survey

2235232650

College of Physicians and Surgeons
Province of Alberta

## Medical Colleague Questionnaire

PAR Physician Achievement Review

Assessed Physician's Name: Dr.

Your Name: Dr.

**Marking Instructions**

Please indicate your answer by filling in the bubbles *like this,* ● **not like** ⊗ or ✓. Thank you!

**How would you describe your professional relationship to this physician (select one)?**

- O Peer (similar practice)
- O Consultant
- O Referring Physician
- O Other (please describe):

**How well do you know this physician (Mark one)?**

O Not at All O Not Well O Somewhat O Well O Very Well

**Interpretation of the Rating Scale**

This form is used by a variety of physicians' colleagues, therefore, not all of the following items may be relevant to you. If any of these items are **NOT** relevant to you, mark these "**Unable to Assess**".

Please rate your colleague on each of the performance statements listed according to the following scale.

**Compared to other medical specialists I know, this one is:**

| | Among the Worst 1 | Bottom Half 2 | Average 3 | Top Half 4 | Among the Best 5 | Unable to Assess UA |
|---|---|---|---|---|---|---|
| 1. Communicates effectively with patients | O | O | O | O | O | O |
| 2. Communicates effectively with patients' families | O | O | O | O | O | O |
| 3. Communicates effectively with other health care professionals | O | O | O | O | O | O |
| 4. Communicates treatment options to patients | O | O | O | O | O | O |
| 5. Within the range of services provided by this physician, he/she performs technical procedures skillfully | O | O | O | O | O | O |
| 6. Selects diagnostic tests appropriately | O | O | O | O | O | O |
| 7. Critically assesses diagnostic information | O | O | O | O | O | O |
| 8. Makes the correct diagnosis following consultation | O | O | O | O | O | O |
| 9. Selects appropriate treatments | O | O | O | O | O | O |
| 10. Maintains quality medical records | O | O | O | O | O | O |
| 11. Handles transfer of care appropriately | O | O | O | O | O | O |
| 12. Provides a clear understanding about who is responsible for the continuing care of patients | O | O | O | O | O | O |
| 13. Recognizes psychosocial aspects of illness | O | O | O | O | O | O |
| 14. Maintains confidentiality of patients and their families | O | O | O | O | O | O |
| 15. Co-ordinates care effectively for patients with other health care professionals and physicians | O | O | O | O | O | O |

*Please turn over*

M

# Appendix B: Medical Colleague Survey

8236232657

| | Compared to other medical specialists I know, this one is: | | | | | |
|---|---|---|---|---|---|---|
| | Among the Worst 1 | Bottom Half 2 | Average 3 | Top Half 4 | Among the Best 5 | Unable to Assess UA |
| 16. Manages patients with complex problems | ○ | ○ | ○ | ○ | ○ | ○ |
| 17. Respects the rights of patients | ○ | ○ | ○ | ○ | ○ | ○ |
| 18. Shows compassion for patients and their families | ○ | ○ | ○ | ○ | ○ | ○ |
| 19. Collaborates with physician colleagues | ○ | ○ | ○ | ○ | ○ | ○ |
| 20. Is involved with professional development | ○ | ○ | ○ | ○ | ○ | ○ |
| 21. Accepts responsibility for own professional actions | ○ | ○ | ○ | ○ | ○ | ○ |
| 22. Manages health care resources efficiently | ○ | ○ | ○ | ○ | ○ | ○ |
| 23. Makes appropriate use of community resources for psychosocial aspects of care | ○ | ○ | ○ | ○ | ○ | ○ |
| 24. Gives priority to urgent requests | ○ | ○ | ○ | ○ | ○ | ○ |
| 25. Handles emergency situations effectively | ○ | ○ | ○ | ○ | ○ | ○ |
| 26. Manages own stress effectively | ○ | ○ | ○ | ○ | ○ | ○ |
| 27. Participates in a system of call to provide care for his/her own patients when unavailable | ○ | ○ | ○ | ○ | ○ | ○ |
| 28. Recognizes his/her own limitations | ○ | ○ | ○ | ○ | ○ | ○ |
| 29. Handles requests for consultation in a timely manner | ○ | ○ | ○ | ○ | ○ | ○ |
| 30. Advises referring physician if referral request is outside the scope of his/her practice | ○ | ○ | ○ | ○ | ○ | ○ |
| 31. Assumes appropriate responsibility for patients | ○ | ○ | ○ | ○ | ○ | ○ |
| 32. Provides timely information to referring physicians about mutual patients | ○ | ○ | ○ | ○ | ○ | ○ |
| 33. Critically evaluates the medical literature to optimize clinical decision making | ○ | ○ | ○ | ○ | ○ | ○ |
| 34. Facilitates the learning of medical colleagues and co-workers | ○ | ○ | ○ | ○ | ○ | ○ |
| 35. Contributes to quality improvement programs and practice guidelines | ○ | ○ | ○ | ○ | ○ | ○ |
| 36. Participates effectively as a member of the health care team | ○ | ○ | ○ | ○ | ○ | ○ |
| 37. Exhibits professional and ethical behavior towards physician colleagues | ○ | ○ | ○ | ○ | ○ | ○ |
| 38. If a member of my own family needed care I would rate this physician | ○ | ○ | ○ | ○ | ○ | ○ |

# Appendix C: Patient Survey

8173160972

College of Physicians and Surgeons
Province of Alberta

## Patient Questionnaire

PAR
Physician Achievement Review

**Gender:** ○ Male ○ Female

**Age:** ○ Less than 1 ○ 19 - 45
○ 1 - 5 ○ 46 - 65
○ 6 - 10 ○ 66 and over
○ 11 - 18

**This questionnaire is being completed by:**
○ Self(patient) ○ Caregiver/parent

**Physician's Name: Dr.**

**Marking Instructions**
Please indicate your answer by filling in the bubbles *like this,* ● **not like** ⊗ **or** ✓. Thank you!

**Interpretation of the Rating Scale**
This form is used by a variety of patients, therefore, not all of the following items may be relevant to you. If any of these items are **NOT** relevant to you, mark these "**Unable to Assess**".

Indicate how much you agree with the statements on the left side of the page using the following scale.

| | Strongly Disagree 1 | Disagree 2 | Neutral 3 | Agree 4 | Strongly Agree 5 | Unable to Assess UA |
|---|---|---|---|---|---|---|
| **Based on my MOST RECENT VISIT, this doctor:** | | | | | | |
| 1. Explained my illness or concern to me clearly | ○ | ○ | ○ | ○ | ○ | ○ |
| 2. Explained my treatment choices or options | ○ | ○ | ○ | ○ | ○ | ○ |
| 3. Explained my follow-up plan to me | ○ | ○ | ○ | ○ | ○ | ○ |
| 4. Told me how and when to take my medicine, if medicine was prescribed | ○ | ○ | ○ | ○ | ○ | ○ |
| 5. Told me of side effects of the medicine, if medicine was prescribed | ○ | ○ | ○ | ○ | ○ | ○ |
| **Based on ALL OF YOUR VISITS to this doctor, how do you feel about this doctor's attitude and behavior towards you? This doctor:** | | | | | | |
| 6. Spends enough time with me | ○ | ○ | ○ | ○ | ○ | ○ |
| 7. Shows interest in my problems | ○ | ○ | ○ | ○ | ○ | ○ |
| 8. Asks details about my personal life, when appropriate | ○ | ○ | ○ | ○ | ○ | ○ |
| 9. Answers my questions well | ○ | ○ | ○ | ○ | ○ | ○ |
| 10. Examines me appropriately for my problems | ○ | ○ | ○ | ○ | ○ | ○ |
| 11. Treats me with respect | ○ | ○ | ○ | ○ | ○ | ○ |
| 12. Helps me with my fears and worries | ○ | ○ | ○ | ○ | ○ | ○ |
| **Rate each statement about this doctor's office. The office:** | | | | | | |
| 13. Is easy to get into (e.g. wheelchair accessible, parking) | ○ | ○ | ○ | ○ | ○ | ○ |
| 14. Has appropriate waiting areas | ○ | ○ | ○ | ○ | ○ | ○ |
| 15. Examining rooms are adequately sized and have adequate equipment | ○ | ○ | ○ | ○ | ○ | ○ |
| 16. Is clean and in good repair | ○ | ○ | ○ | ○ | ○ | ○ |
| 17. Provides adequate privacy | ○ | ○ | ○ | ○ | ○ | ○ |

*Please turn over*

M Page 1 of 2

# Appendix C: Patient Survey

2611160973

| | Strongly Disagree 1 | Disagree 2 | Neutral 3 | Agree 4 | Strongly Agree 5 | Unable to Assess UA |
|---|---|---|---|---|---|---|
| **How do you feel this doctor runs his or her practice?** | | | | | | |
| **Telephone:** | | | | | | |
| 18. I can reach the office by phone during the day | ○ | ○ | ○ | ○ | ○ | ○ |
| 19. I receive an appropriate explanation if my appointment is delayed | ○ | ○ | ○ | ○ | ○ | ○ |
| 20. My messages are returned | ○ | ○ | ○ | ○ | ○ | ○ |
| **The Staff:** | | | | | | |
| 21. Are helpful and pleasant | ○ | ○ | ○ | ○ | ○ | ○ |
| 22. Are respectful of patients | ○ | ○ | ○ | ○ | ○ | ○ |
| 23. Behave in a professional manner | ○ | ○ | ○ | ○ | ○ | ○ |
| 24. Work well with the doctor | ○ | ○ | ○ | ○ | ○ | ○ |
| 25. Prevent patients from hearing confidential information about other patients | ○ | ○ | ○ | ○ | ○ | ○ |
| **Office Practices:** | | | | | | |
| 26. In an emergency situation this doctor's office provides me with clear instructions on what I am to do | ○ | ○ | ○ | ○ | ○ | ○ |
| 27. This doctor provides reports to my family doctor | ○ | ○ | ○ | ○ | ○ | ○ |
| 28. When asked, this doctor provides insurance and medico legal reports in a timely manner | ○ | ○ | ○ | ○ | ○ | ○ |
| 29. When asked, this doctor provides reports, files or copies of letters in a timely manner | ○ | ○ | ○ | ○ | ○ | ○ |
| 30. This doctor arranges appointments with other specialists when necessary | ○ | ○ | ○ | ○ | ○ | ○ |
| 31. This doctor's office follows-up on serious problems | ○ | ○ | ○ | ○ | ○ | ○ |
| 32. I am told what to do if my problems do not get better | ○ | ○ | ○ | ○ | ○ | ○ |
| **General:** | | | | | | |
| 33. I am asked about prescription and non-prescription medicine I may be taking | ○ | ○ | ○ | ○ | ○ | ○ |
| 34. This doctor talks to me about preventative care (e.g. quitting smoking, weight control, sleeping, alcohol, exercise, etc.) | ○ | ○ | ○ | ○ | ○ | ○ |
| 35. This doctor has good written health information | ○ | ○ | ○ | ○ | ○ | ○ |
| 36. This doctor refers me to appropriate educational resources (i.e., web sites, brochures, patient support groups, books) | ○ | ○ | ○ | ○ | ○ | ○ |
| 37. I would go back to this doctor | ○ | ○ | ○ | ○ | ○ | ○ |
| 38. I would send a friend to this doctor | ○ | ○ | ○ | ○ | ○ | ○ |
| 39. This doctor presents him/herself in a professional manner | ○ | ○ | ○ | ○ | ○ | ○ |
| 40. I was helped by this doctor | ○ | ○ | ○ | ○ | ○ | ○ |

# Appendix D: Self Survey

8396116396

College of Physicians and Surgeons
Province of Alberta

## Self Assessment Questionnaire

PAR
Physician Achievement Review

Name: Dr.

**Marking Instructions**
Please indicate your answer by filling in the bubbles *like this,* ● **not like** ⊗ **or** ✓. Thank you!

**Interpretation of the Rating Scale**
This form is used by a variety of medical specialists, therefore, not all of the following items may be relevant to you. If any of these items are **NOT** relevant to you, mark these "**Unable to Assess**".

The following statements describe physician behaviours. Please rate yourself on each of the performance statements listed using the scale to the right.

**Compared to other medical specialists you know, please rate your performance for each statement:**

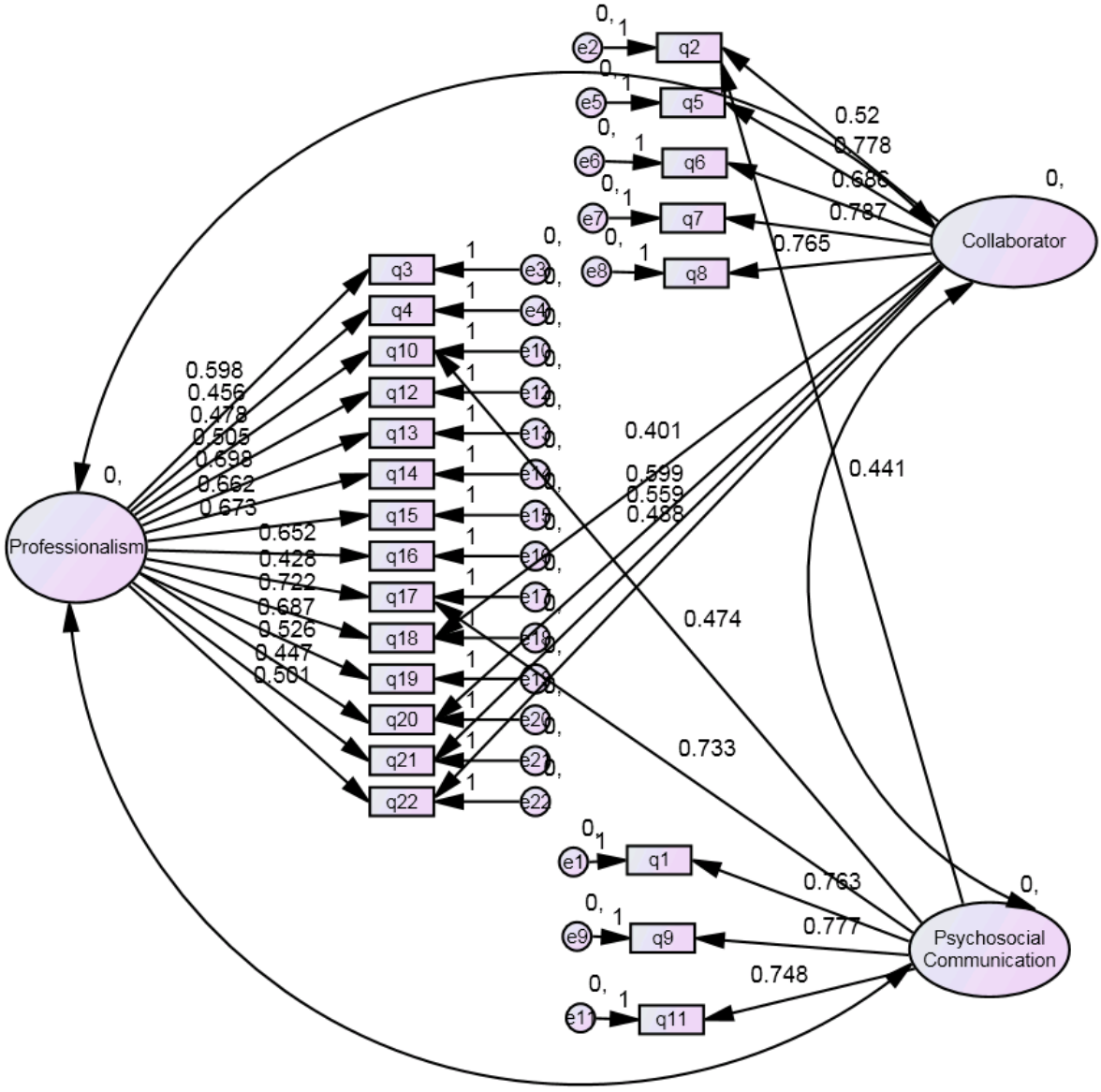
| | | Among the Worst 1 | Bottom Half 2 | Average 3 | Top Half 4 | Among the Best 5 | Unable to Assess UA |
|---|---|---|---|---|---|---|---|
| 1. | I communicate effectively with patients | ○ | ○ | ○ | ○ | ○ | ○ |
| 2. | I communicate effectively with patients' families | ○ | ○ | ○ | ○ | ○ | ○ |
| 3. | I communicate effectively with other health care professionals | ○ | ○ | ○ | ○ | ○ | ○ |
| 4. | I communicate treatment options to patients | ○ | ○ | ○ | ○ | ○ | ○ |
| 5. | Within the range of services provided by me, I perform technical procedures skillfully | ○ | ○ | ○ | ○ | ○ | ○ |
| 6. | I select diagnostic tests appropriately | ○ | ○ | ○ | ○ | ○ | ○ |
| 7. | I critically assess diagnostic information | ○ | ○ | ○ | ○ | ○ | ○ |
| 8. | I make the correct diagnosis following consultation | ○ | ○ | ○ | ○ | ○ | ○ |
| 9. | I select appropriate treatments | ○ | ○ | ○ | ○ | ○ | ○ |
| 10. | I maintain quality medical records | ○ | ○ | ○ | ○ | ○ | ○ |
| 11. | I handle transfer of care appropriately | ○ | ○ | ○ | ○ | ○ | ○ |
| 12. | I provide a clear understanding about who is responsible for the continuing care of patients | ○ | ○ | ○ | ○ | ○ | ○ |
| 13. | I recognize psychosocial aspects of illness | ○ | ○ | ○ | ○ | ○ | ○ |
| 14. | I maintain confidentiality of patients and their families | ○ | ○ | ○ | ○ | ○ | ○ |
| 15. | I co-ordinate care effectively for patients with other health care professionals and physicians | ○ | ○ | ○ | ○ | ○ | ○ |
| 16. | I manage patients with complex problems | ○ | ○ | ○ | ○ | ○ | ○ |

*Please turn over*

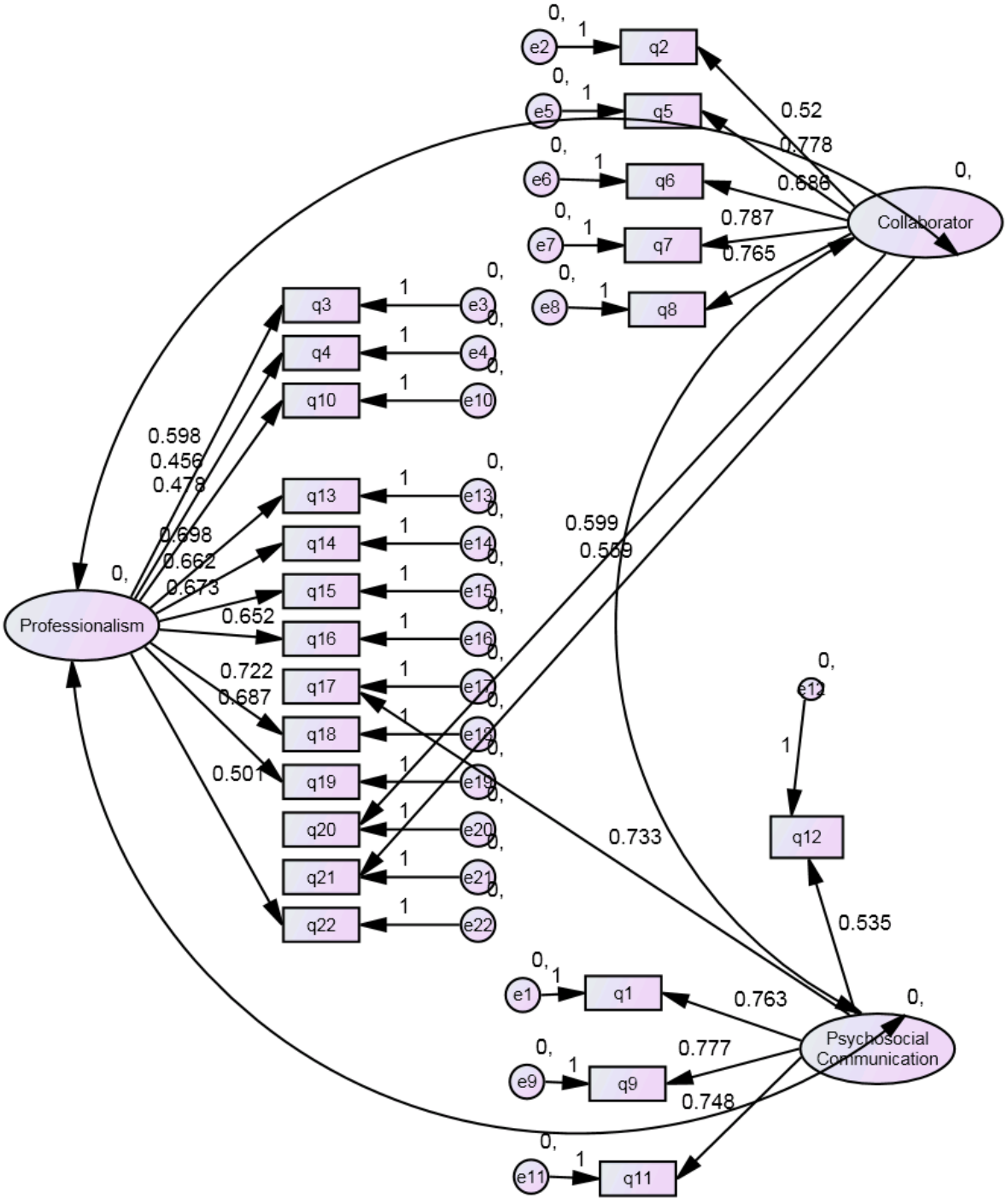
# Appendix D: Self Survey

| | Compared to other medical specialists you know, please rate your performance for each statement: | | | | | |
|---|---|---|---|---|---|---|
| | Among the Worst **1** | Bottom Half **2** | Average **3** | Top Half **4** | Among the Best **5** | Unable to Assess **UA** |
| 17. I respect the rights of patients | ○ | ○ | ○ | ○ | ○ | ○ |
| 18. I show compassion for patients and their families | ○ | ○ | ○ | ○ | ○ | ○ |
| 19. I collaborate with physician colleagues | ○ | ○ | ○ | ○ | ○ | ○ |
| 20. I am involved with professional development | ○ | ○ | ○ | ○ | ○ | ○ |
| 21. I accept responsibility for my professional actions | ○ | ○ | ○ | ○ | ○ | ○ |
| 22. I manage health care resources efficiently | ○ | ○ | ○ | ○ | ○ | ○ |
| 23. I make appropriate use of community resources for psychosocial aspects of care | ○ | ○ | ○ | ○ | ○ | ○ |
| 24. I give priority to urgent requests | ○ | ○ | ○ | ○ | ○ | ○ |
| 25. I handle emergency situations effectively | ○ | ○ | ○ | ○ | ○ | ○ |
| 26. I manage my stress effectively | ○ | ○ | ○ | ○ | ○ | ○ |
| 27. I participate in a system of call to provide care for my patients when I am unavailable | ○ | ○ | ○ | ○ | ○ | ○ |
| 28. I recognize my limitations | ○ | ○ | ○ | ○ | ○ | ○ |
| 29. I handle requests for consultation in a timely manner | ○ | ○ | ○ | ○ | ○ | ○ |
| 30. I advise referring physician if referral request is outside the scope of my practice | ○ | ○ | ○ | ○ | ○ | ○ |
| 31. I assume appropriate responsibility for patients | ○ | ○ | ○ | ○ | ○ | ○ |
| 32. I provide timely information to referring physicians about mutual patients | ○ | ○ | ○ | ○ | ○ | ○ |
| 33. I critically evaluate the medical literature to optimize clinical decision making | ○ | ○ | ○ | ○ | ○ | ○ |
| 34. I facilitate the learning of medical colleagues and co-workers | ○ | ○ | ○ | ○ | ○ | ○ |
| 35. I contribute to quality improvement programs and practice guidelines | ○ | ○ | ○ | ○ | ○ | ○ |
| 36. I participate effectively as a member of the health care team | ○ | ○ | ○ | ○ | ○ | ○ |
| 37. I exhibit professional and ethical behavior towards physician colleagues | ○ | ○ | ○ | ○ | ○ | ○ |

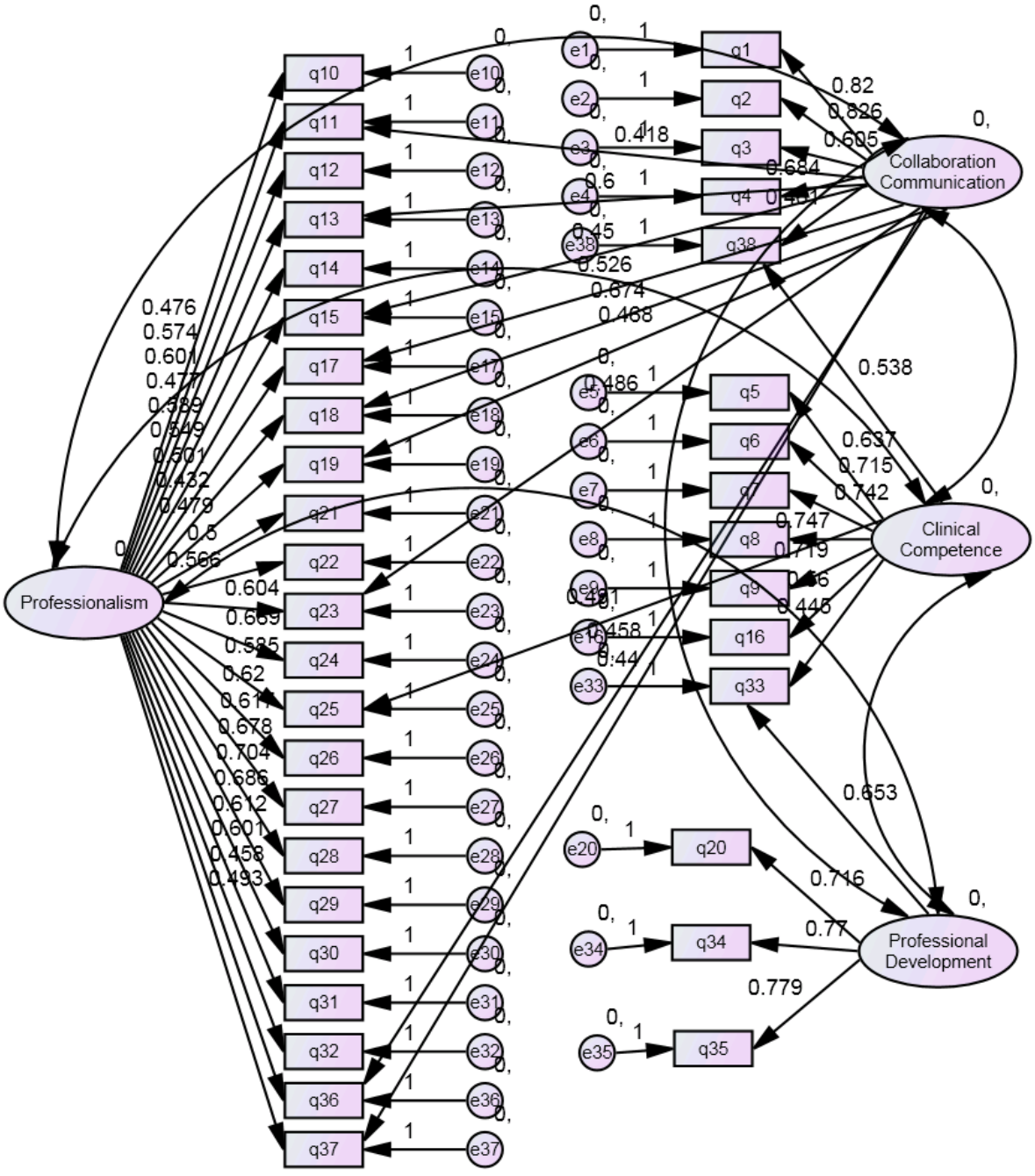
**Appendix E** CFA Initial Model Diagram, Coworker Survey



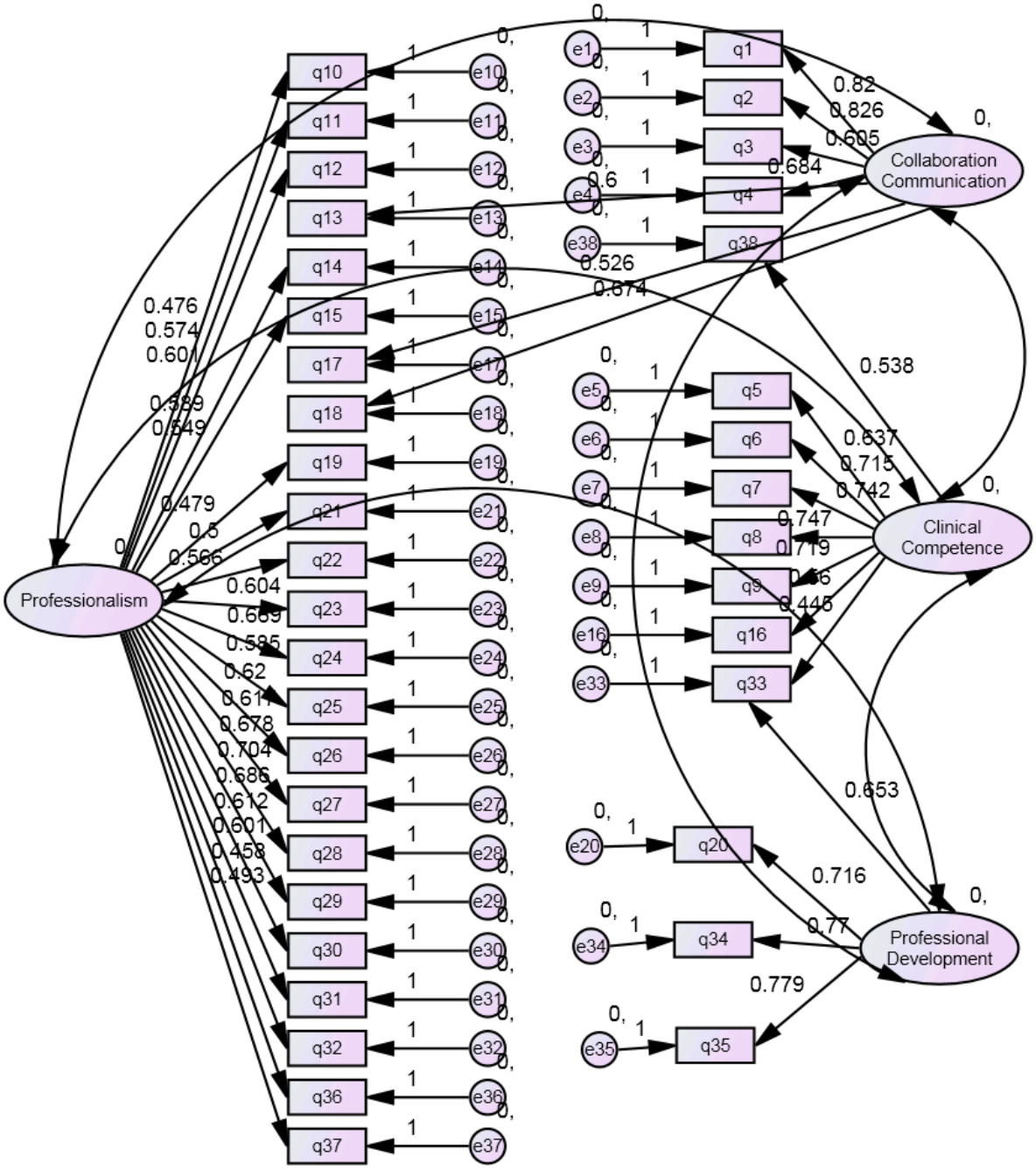
**Appendix F** CFA Rival Model Diagram, Coworker Survey



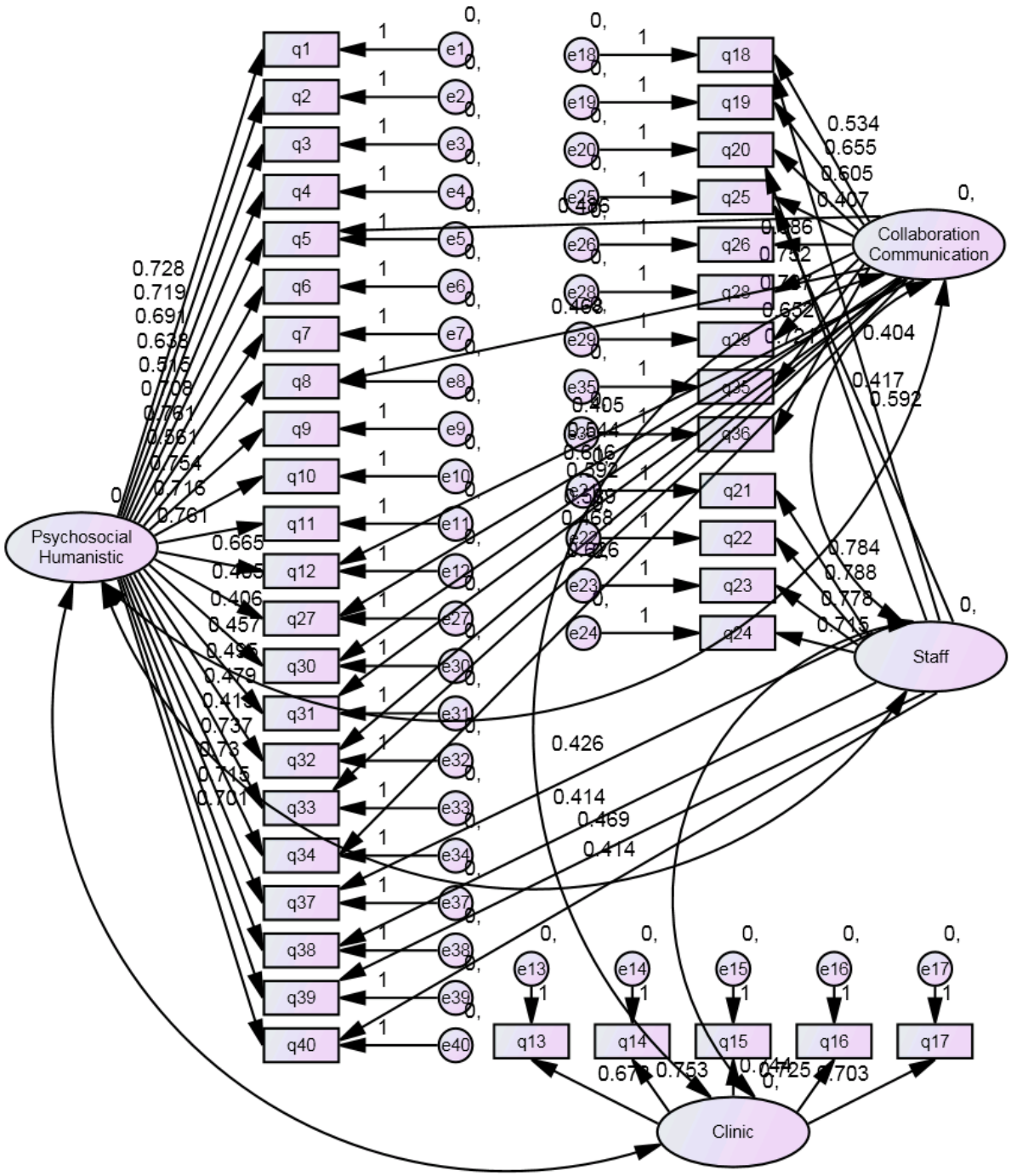
**Appendix G** CFA Initial Model Diagram, Medical Colleague Survey



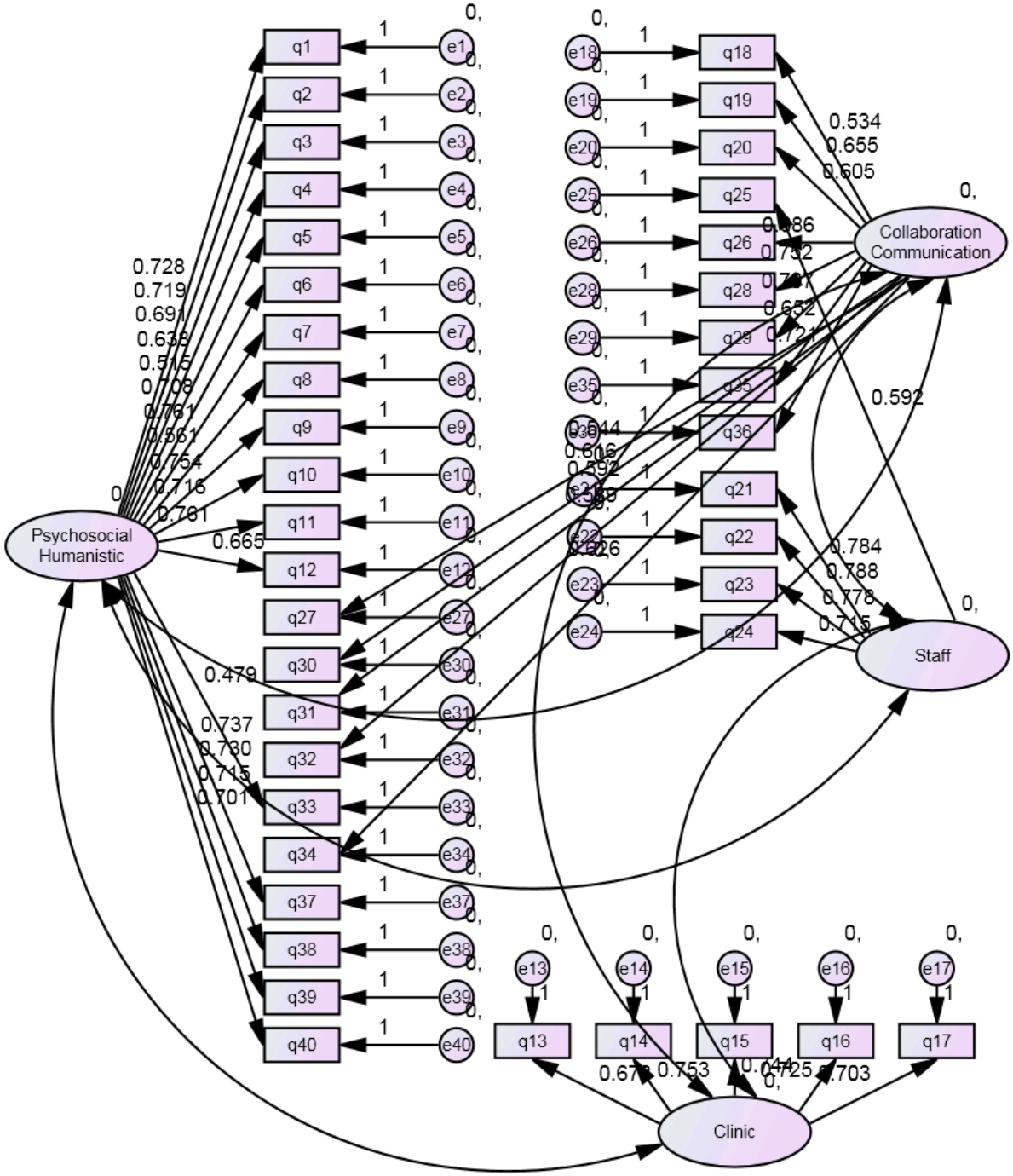
**Appendix H** CFA Rival Model Diagram, Medical Colleague Survey



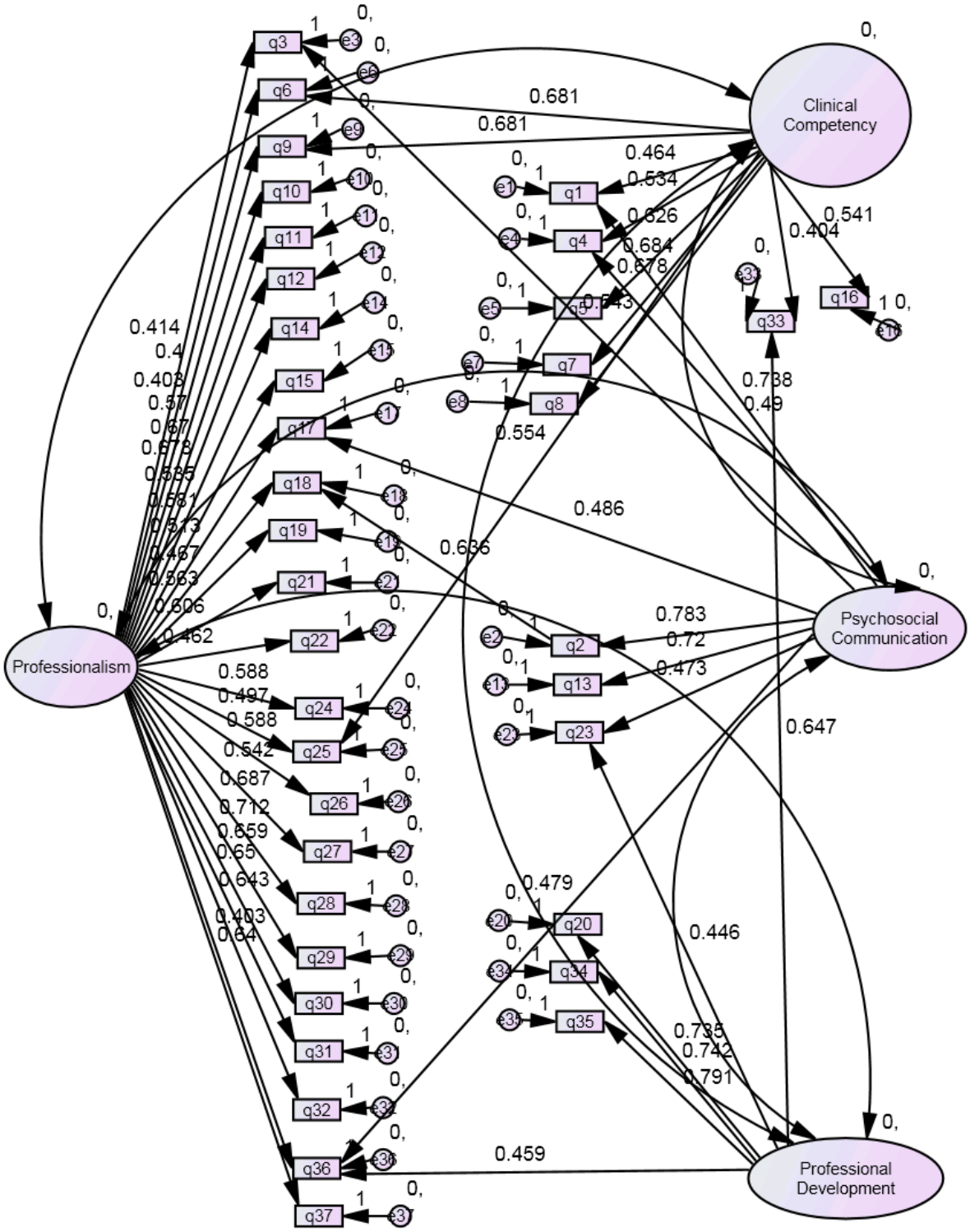
**Appendix I** CFA Initial Model Diagram, Patient Survey



**Appendix J** CFA Rival Model Diagram, Patient Survey



**Appendix K** CFA Initial Model Diagram, Self Survey



**Appendix L** CFA Rival Model Diagram, Self Survey