DEVELOPMENT AND ANALYSIS OF A SERVICE BEHAVIOR INSTRUMENT FOR USE IN THE HEALTHCARE SETTING

A dissertation submitted to the
Kent State University
College of Education, Health, and Human Services
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

By
Edward C. Bolden, III
August, 2015
The purpose of this study was to develop and evaluate the psychometric properties of an instrument that assesses the service behaviors of employees at a large hospital in the Midwest. There has been an increasing focus on the service quality in healthcare settings as legislation for reform is drawing attention to quality and value for healthcare consumers (Kennedy, Caselli, & Berry, 2011). Hospitals and other healthcare settings are increasingly concerned about patient experience and perceptions of service quality in the competitive healthcare marketplace (Scotti, Harmon, & Beson, 2007). Hospitals are continually seeking to provide services that differentiate them from others and enhance the patient experience.

The pilot testing phase consisted of 165 participants who were new-hire employees that completed the instrument as part of their initial hospital orientation to gather information on their perceptions of service behavior delivery. Following information provided throughout the pilot testing phase, the instrument was revised and administered to 452 hospital employees over a three month span. Data collected during both pilot testing and the full study phase were used to guide changes and assess the psychometric properties of the instrument.
ACKNOWLEDGEMENTS

I would first like to thank my committee for the tremendous amount of time and insight they have provided me along the way. Their feedback and support was critical to my perseverance with this project. I would particularly like to thank Dr. Jian Li for her guidance and dedication over the past year and a half. Additionally, I would like to recognize Dr. Jason Schenker and Dr. Aryn Karpinski, who have shared with me their passion for evaluation and measurement as instructors throughout my coursework. They have challenged me on so many levels, and for that I am truly grateful.

Many other individuals have been immensely supportive of my professional journey, including Dr. Deborah Shama-Davis and Pamela Freeman. I owe them both so much credit for allowing me access to this project and for providing me additional support throughout the entire process. They have been fantastic leaders, mentors, and colleagues, and I am sincerely grateful for the opportunities they have presented to me over the years. I truly cannot thank them both enough for all they have done for me.

I would like to dedicate this in loving memory of my mother, Karen Marie, who is missed every day of my life. I would also like to dedicate this to my father, who has always led by example of being the strongest, hardest working man I know. I want to thank him for everything he has done to challenge, motivate, and push me to be the best person I could ever be. I would truly not be here without the support and love he has shown for me. A very special thank you to my sisters, Cynthia and Jaclyn, who have motivated me in their own ways, and with loving thanks to the rest of my entire family.
for your enduring support, including those that are no longer with us. Thank you all for making me who I am today.

Another very special thank you to Jayne, who has always shown an interest in what I have been working on and who is always willing to talk research with me. I am so excited she has been able to uncover her own passion for research and has taken initiative sharing that with her colleagues. Thank you for always encouraging me to discuss my work and for following your interest in research as well.

Last but certainly not least, I would like to thank my wife, Heather. She has provided me with her patience, reassurance, and support throughout this entire journey, helping me every step along the way. This process has at times been challenging, frustrating, and overall time-consuming, and I appreciate her willingness and understanding. I am so grateful to be sharing this with someone so special and someone that has believed in me since this was an idea I had many years ago. I want to thank her for the continuous support she has provided me, her love, and for our beautiful blessing, Kali Grace.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>iv</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>ix</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>x</td>
</tr>
<tr>
<td>CHAPTER</td>
<td></td>
</tr>
<tr>
<td>I. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>Overview</td>
<td>1</td>
</tr>
<tr>
<td>Rationale and Purpose</td>
<td>2</td>
</tr>
<tr>
<td>Goals and Research Questions</td>
<td>4</td>
</tr>
<tr>
<td>Significance and Limitations</td>
<td>5</td>
</tr>
<tr>
<td>II. REVIEW OF THE LITERATURE</td>
<td>8</td>
</tr>
<tr>
<td>A History of Service Quality Measurement</td>
<td>8</td>
</tr>
<tr>
<td>Service Quality</td>
<td>8</td>
</tr>
<tr>
<td>Existing Measures of Service Quality in General Settings</td>
<td>12</td>
</tr>
<tr>
<td>Service Quality in the Healthcare Setting</td>
<td>20</td>
</tr>
<tr>
<td>The Instrument Development and Refinement Process</td>
<td>30</td>
</tr>
<tr>
<td>Item Development and Pilot Testing</td>
<td>32</td>
</tr>
<tr>
<td>Instrument Refinement and Confirmatory Analyses</td>
<td>33</td>
</tr>
<tr>
<td>Methods and Models Employed in the Psychometric Property Examination</td>
<td>33</td>
</tr>
<tr>
<td>Item Analysis</td>
<td>34</td>
</tr>
<tr>
<td>Validity</td>
<td>34</td>
</tr>
<tr>
<td>Reliability</td>
<td>35</td>
</tr>
<tr>
<td>Exploratory Factor Analysis</td>
<td>36</td>
</tr>
<tr>
<td>Confirmatory Factor Analysis</td>
<td>38</td>
</tr>
<tr>
<td>Rasch Analysis</td>
<td>40</td>
</tr>
<tr>
<td>III. METHODOLOGY</td>
<td>44</td>
</tr>
<tr>
<td>Part I: Pilot Testing Phase</td>
<td>44</td>
</tr>
<tr>
<td>Participants</td>
<td>44</td>
</tr>
<tr>
<td>Measure</td>
<td>45</td>
</tr>
<tr>
<td>Item development</td>
<td>45</td>
</tr>
<tr>
<td>The pilot instrument</td>
<td>51</td>
</tr>
</tbody>
</table>
IV. ANALYSIS OF RESULTS

Part I: Pilot Testing Phase

Sample ................................................................. 63
Item Analysis .......................................................... 65
Correlations .............................................................. 72
Exploratory Factor Analysis of the 46 Rating-Scale Items .......... 76
Refinement ............................................................... 80
Exploratory Factor Analysis of the 30 Retained Items .............. 87
Vignettes ................................................................ 91
Finalizing the Full-Scale Instrument .................................. 92

Part II: Full-Scale Administration

Sample ................................................................ 93
Item Analysis of 30 "Patients/Visitors" Items ......................... 95
Item Analysis of 9 "Colleagues" Items ............................... 97
Vignettes ................................................................ 99
Data Reduction .......................................................... 100
Principal Components Analysis ....................................... 101
Confirmatory Factor Analysis ......................................... 108
Rasch Analysis ............................................................ 122
Scale and item statistics .................................................. 123
Person statistics ......................................................... 123
Item statistics ............................................................. 129
Person ability and item difficulty ..................................... 132
Category structure and usage ......................................... 134
Underlying factor structure and dimensionality .................. 138
Differential item functioning .......................................... 144
DIF analysis of gender groups ....................................... 145
DIF analysis of direct patient care contact groups .............. 145
<table>
<thead>
<tr>
<th>V. DISCUSSION, LIMITATIONS, AND RECOMMENDATIONS</th>
<th>153</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discussion ................................................</td>
<td>153</td>
</tr>
<tr>
<td>Pilot Testing Phase .......................................</td>
<td>153</td>
</tr>
<tr>
<td>Data collection and sample ................................</td>
<td>154</td>
</tr>
<tr>
<td>Analyses ....................................................</td>
<td>155</td>
</tr>
<tr>
<td>Refinement ..................................................</td>
<td>156</td>
</tr>
<tr>
<td>Full-Scale Administration ..................................</td>
<td>156</td>
</tr>
<tr>
<td>Data collection and sample ................................</td>
<td>156</td>
</tr>
<tr>
<td>Analyses ....................................................</td>
<td>157</td>
</tr>
<tr>
<td>Item analysis ...............................................</td>
<td>158</td>
</tr>
<tr>
<td>Principal components analysis ............................</td>
<td>158</td>
</tr>
<tr>
<td>Confirmatory factor analysis ..............................</td>
<td>159</td>
</tr>
<tr>
<td>Rasch analysis ...............................................</td>
<td>159</td>
</tr>
<tr>
<td>Implications ..................................................</td>
<td>161</td>
</tr>
<tr>
<td>Limitations ..................................................</td>
<td>163</td>
</tr>
<tr>
<td>Recommendations .............................................</td>
<td>165</td>
</tr>
</tbody>
</table>

| APPENDICES .................................................. | 171 |
| APPENDIX A. PILOT INSTRUMENT ............................ | 172 |
| APPENDIX B. FULL-SCALE INSTRUMENT ........................ | 179 |
| APPENDIX C. RHO CORRELATIONS BETWEEN PATIENT/VISITOR AND COLLEAGUE RATINGS | 184 |

<p>| REFERENCES ................................................... | 187 |</p>
<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. The initial model</td>
<td>109</td>
</tr>
<tr>
<td>2. The final model</td>
<td>116</td>
</tr>
<tr>
<td>3. The map of persons and items for the service behavior instrument</td>
<td>133</td>
</tr>
<tr>
<td>4. Category probability curves</td>
<td>136</td>
</tr>
<tr>
<td>5. Category probability curves updated with the collapsed categories</td>
<td>138</td>
</tr>
<tr>
<td>6. Scree plot of standardized residual variance</td>
<td>140</td>
</tr>
<tr>
<td>7. Component factor plot</td>
<td>143</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table                                      Page

1. Expected Service Behaviors and Number of Corresponding Items on Pilot Instrument ................................................................. 48

2. Pilot Instrument Responses by Completion Percentage ........................................ 64

3. Cross-Tabulation between Gender and Instrument Completion ............................. 65

4. Descriptive Statistics and Alpha if Item Deleted for Items on the Pilot Instrument .................................................................................................................. 68

5. Item-Total Correlations for Adapted Items Rated for Patients/Visitors and Colleagues ........................................................................................................ 75

6. Internal Consistency within Item Groupings by Behavioral Indicators .................. 85

7. Results of the Exploratory Factor Analysis with Items Removed .......................... 89

8. Percent of Correct Responses on Vignette Items .................................................. 91

9. Cross-Tabulation between Gender and Direct Patient Care Contact on the Full-Scale Administration .................................................................................. 94

10. Descriptive Statistics and Alpha if Item Removed for Patient/Visitor Items on the Full-Scale Instrument .............................................................. 96

11. Descriptive Statistics and Alpha if Item Removed for Colleague Items on the Full-Scale Instrument ................................................................. 98

12. Percent of Correct Responses on Vignette Items ................................................. 99
<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>13. Results of the Principal Components Analysis with Varimax Rotation</td>
<td>106</td>
</tr>
<tr>
<td>14. Initial Confirmatory Factor Analysis Results: Fit Indices</td>
<td>111</td>
</tr>
<tr>
<td>15. Modified Model Confirmatory Factor Analysis Results: Fit Indices</td>
<td>112</td>
</tr>
<tr>
<td>16. Confirmatory Factor Analysis Results: Fit Indices for Three Models</td>
<td>114</td>
</tr>
<tr>
<td>17. Final Confirmatory Factor Analysis Results: Fit Indices for Four Models</td>
<td>118</td>
</tr>
<tr>
<td>18. Correlation Matrix for Four Factors</td>
<td>119</td>
</tr>
<tr>
<td>19. Summary of 256 Measured (Non-Extreme) Persons (30 Measured Items)</td>
<td>124</td>
</tr>
<tr>
<td>20. Person Statistics: Misfit Order (30 Measured Items)</td>
<td>127</td>
</tr>
<tr>
<td>21. Summary of 222 Measured (Non-Extreme) Persons (30 Measured Items)</td>
<td>128</td>
</tr>
<tr>
<td>22. Summary of 30 Measured Items (222 Measured Persons)</td>
<td>130</td>
</tr>
<tr>
<td>23. Item Statistics: Misfit Order (222 Measured Persons)</td>
<td>131</td>
</tr>
<tr>
<td>24. Category Frequency and Threshold Values</td>
<td>135</td>
</tr>
<tr>
<td>25. Category Frequency and Threshold Values with Collapsed Categories</td>
<td>137</td>
</tr>
<tr>
<td>26. Standardized Residual Variance (in Eigenvalue units)</td>
<td>141</td>
</tr>
<tr>
<td>27. Standardized Residual Loadings for Items in Contrast 1 (Sorted by Loading)</td>
<td>142</td>
</tr>
<tr>
<td>28. DIF Class Specification for Patient Care Contact Groups</td>
<td>147</td>
</tr>
</tbody>
</table>
CHAPTER I

INTRODUCTION

In this chapter, a brief overview of the existing literature is offered to provide background on the rationale and purpose of the research. The unique contribution of the project, its anticipated significance, and potential limitations are also discussed. The chapter concludes with a discussion of the goals of the research and a presentation of the research questions, which will serve to guide the investigation.

Overview

There has been an increasing focus on the service quality in healthcare settings as legislation for reform is drawing attention to quality and value for healthcare consumers (Kennedy, Caselli, & Berry, 2011). Hospitals and other healthcare settings are becoming increasingly concerned about patient experience and perceptions of service quality in the competitive healthcare marketplace (Scotti, Harmon, & Beson, 2007). In order to best compete for patients, hospitals are continually seeking to provide quality services that differentiate them from others and enhance the patient experience. Further, measuring patient experience to drive quality improvement and patient choice is now mandatory in many places (Black & Jenkinson, 2009). These services extend beyond practical application of medical skills and have become more focused on specific service behaviors that foster a positive patient experience; for example, showing empathy to patients.

Many of these service behaviors are universal, applicable to all areas of service encounters in which the customer is engaged (Parasuraman, Zeithaml, & Berry, 1988).
These service behaviors are related to the success of the business in which they are applied, where businesses that deliver high quality service are able to be positively differentiated from those of less quality (Rudie & Wansley, 1985; Thompson, DeSouza, & Gale, 1985). Beyond differentiation, there is a demonstrable connection between service behaviors or the quality of the service encounter and corresponding customer satisfaction (Winsted, 2000). Customer satisfaction and perceptions of the service encounter are essential to acknowledge as they relate to the success of the business.

While there are several assessments of service quality available, such as the SERVQUAL (Parasuraman, Zeithaml, & Berry, 1988) scale, the dimensions and meaning of identified service quality dimensions remains unspecified (Boulding, Kalra, Staelin, & Zeithaml, 1993). The healthcare system is itself different from many other service industries, to which application of techniques or organizational practices may not translate effectively (Scotti, Harmon, & Beson, 2007). There seem to be few instruments assessing service quality through operational, objective behaviors that apply directly to the healthcare setting, which makes assessing specific behaviors and measuring behavior change difficult. Many of those that have been used have not been subject to rigorous evaluation of their validity, reliability, and other psychometric properties that allow them to be objective measures of service behaviors representative of the healthcare field.

**Rationale and Purpose**

The rationale for developing this instrument is derived from the increasing need for healthcare systems to evaluate their employee-patient service encounters. Healthcare
settings need to be proactive in evaluating their quality of service through employee behaviors as the legislation for reform is promoting competition within the healthcare marketplace. Specifically, there exists a need to directly, accurately, and reliably assess employee behaviors that represent strong service quality. Given the increased attention directed at service quality and patient experience, hospitals and other healthcare settings need to be able to objectively monitor and evaluate their employees’ service behaviors to assess their own strengths and weaknesses relative to service quality.

When looking within the healthcare discipline for a benchmark of service quality, many employee evaluations lack specificity in addressing service behaviors, instead attending to behaviors of professionalism. Existing measures of service quality lack the focus on behaviors specific to healthcare settings, while those that are healthcare specific have yet be subject to analyses that require demonstration of sound psychometric properties. There remains a need for an instrument to evaluate behaviors specific to service quality within the healthcare setting, which can demonstrate adequate scale properties rarely assessed in the field. The current instrument will be subject to analyses of validity, reliability, and underlying factor structure, as well as an analysis using Item Response Theory principles to demonstrate the measure has sound psychometric properties and evaluate how it captures the construct of service quality.

The purpose of this study was twofold: First, this research sought to develop an instrument that adequately addresses the service behaviors that are important to the hospital setting used in this research. The intention was to provide uses within the
industry with a way to assess their own strengths and weaknesses relative to quality of
service delivery. Secondly, to evaluate the psychometric properties and underlying factor
structure of the instrument to provide evidence of validity and reliability, as well as
present a refined, succinct measure of service behaviors in the healthcare setting. The
instrument was developed beginning with the generation of items that adequately address
the behaviors selected by the focus institution. Pilot testing took place prior to full-scale
administration as a means of producing informative feedback regarding the items and
assist in refinement of the measure. Psychometric properties of both the pilot instrument
and full scale measure were evaluated to provide insight into reliability of the instrument
and potential for generalizability to other healthcare settings.

**Goals and Research Questions**

The goals of the current research follow the instrument development process. The
goal of the pilot testing phase were to gather information and feedback relevant to
instrument design and item development to dictate refinement, as well as demonstrate
concurrent validation with other assessments of service quality. The revised instrument
was expected to demonstrate sound psychometric properties and provide valid, reliable
results that contribute meaningful data and feedback to the hospital. The overarching goal
of the research was to generate an instrument that can be applied in the healthcare setting
to assess employee performance of target service behaviors. Ultimately, the instrument
should meet the needs of the hospital used in this research—as well as other healthcare
settings—in providing feedback regarding their employees’ use of service behaviors and their behavioral impact on patient experience.

While the goals of the current research are to develop a psychometrically sound instrument that demonstrates adequate reliability and validity, the research questions center on evaluating those qualities of the instrument. Specifically, the research will address four questions specific to the instrument through the pilot and final testing phases:

1. Does the newly-developed instrument meet acceptable criteria in terms of reliability and validity, while adequately addressing target service behaviors of the current healthcare setting?
2. Does the service behavior instrument demonstrate psychometric properties that meet the tenets of the Rasch measurement approach, in terms of reliability, separation, and item fit?
3. How is the construct of service quality captured by the instrument, as evidenced by underlying dimensions identified in the Rasch factor analysis?
4. Does the underlying factor structure of the instrument agree with factor structures in existing assessments of service behaviors?

**Significance and Limitations**

The significance of the current investigation is anticipated to extend beyond the single setting utilized in the research and apply to a variety of other healthcare settings. As aforementioned, the purpose of this research is to generate an instrument with sound
psychometric properties that appropriately assesses employee service behaviors in the healthcare setting. The unique contribution provided by this instrument is expected to be not only its design specific to the needs of healthcare providers, but also the ability for the instrument to meet the criteria of a sound, valid, and reliable measure. Instruments that have previously been utilized in the healthcare setting have rarely been subject to such a strict evaluation of measurement quality. It is expected that the instrument developed in this investigation will meet stringent psychometric criteria, enhancing its utility in similar settings and with other populations.

There are several anticipated limitations to the current research, including the identification of the target behaviors, the use of self-report data collection, and other threats to internal and external validity. Primarily, the specific service behaviors will be identified by the hospital setting where the instrument will be administered. The determination of key components of the service encounter was not explored in the current investigation. Instead, the behaviors were identified by the research site and served as the impetus for item development. Thus, there may be subjectivity inherent in what the research site believes to be essential service behaviors and may not adequately reflect all aspects of the service encounter in the healthcare setting. Similarly, the behaviors acknowledged by the current setting may not generalize equally to other healthcare settings. To address this limitation, items from existing service quality measures will be included in the pilot instrument as a means of building evidence of concurrent validity with more established measures of service behavior.
Another potential limitation to the current investigation is the reliance on self-report as the primary means of data collection. Individuals may provide responses that are not accurate, but instead introduce bias in the data to appear more favorably or socially desirable (Shadish, Cook, & Campbell, 2002). This may occur here, where employees are asked to provide self-ratings on what the employer has identified as critical components of patient interaction. In an attempt to control for such participant effects, data will be collected in a manner in which respondents will not be able to be identified. Individuals will not have their responses traced or tied to their employee record in any way. In the event responses cannot be linked to individuals, respondents tend to provide more unbiased self-reports, assisting in the validity of the results (Kazdin, 2003). Other potential limitations to the current research will be discussed as they arise following the presentation and interpretation of the results.

Generalizability of the findings is expected to be reasonable given the limitations, due in part to the large, diverse sample size that is expected to participate. The prospective subject pool is a diverse, heterogeneous group of healthcare workers that is likely representative of other healthcare settings as well. Due to the location and size of the hospital used in the data collection, the diversity of respondents is expected to be similar to other healthcare sites. The participant recruitment and data collection is inclusive of most employees within the hospital, so data from various positions and demographic representations is anticipated. Specific details about the population and generalizability of the results will be presented in subsequent sections.
CHAPTER II
REVIEW OF THE LITERATURE

The review of the literature begins with a presentation of the history of service quality measurement before providing the strengths and limitations of existing instruments. An introduction to service quality and measurement specific to the healthcare setting is also provided. Next, a research framework for the instrument development process is presented and discussed as it applies to the current context. The chapter concludes with a review of the analytical models that will be employed in the data analysis phase.

A History of Service Quality Measurement

The first section presents an overview of the construct of service quality. A review of measures of service quality is presented after, beginning with instruments designed for use in all sectors. The last section provides a review of service quality measurement specific to the healthcare domain.

Service Quality

Service quality became popular in the marketing and business literature in the early 1980s (Parasuraman, Zeithaml, & Berry, 1985), as the rising consumer trend of that decade was the search for quality in products and services (Rabin, 1983). At that time, there was a limited amount of research targeting consumer evaluations of service quality (e.g., Gronroos, 1982; Lewis & Booms, 1983; Sasser, Olsen, & Wyckoff, 1978). Over the years, consumers have continually developed a demand for high quality products and
services. What the early research suggested was that service quality is more difficult to evaluate than quality of the goods, from the consumer’s perspective (Lewis & Booms, 1983). Also, service quality perceptions involve the evaluation of the process of service delivery, comparing consumer expectations with actual service performance (Sasser, Olsen, & Wyckoff, 1978). The uncertainty of consumer expectations and how they evaluate the service encounter makes it challenging for service providers to influence the consumer’s perceptions on quality in the service encounter.

The early research on attitudes provided evidence that consumers enter the service encounter with a set of expectations, and quality is perceived in terms of how the consumer’s expectations are met (Gronroos, 1982). Satisfaction with services is directly related to the consumer’s expectations, where consumers are satisfied when their expectations are met (Smith & Houston, 1982). Customers approach the service encounter with expectations over what should happen, and compare the encounter with their ideal expectation (Zeithaml, Berry, & Parasuraman, 1991). These expectations are phrased in terms of quality, of both the outcome of the service and the process. The quality of the service the consumer actually receives is distinct from the quality of the manner in which the service is delivered (Gronroos, 1982). Consumers carry with them expectations of the service as well as the process of service delivery, which both serve as factors of service quality perceptions.

Consumer expectations are directly related to their consumption experiences, where higher expectations are tied to consumer behaviors (Sellers, 1990). Those with high service expectations that have been met are more likely to continue a productive
relationship with that business, at least in the retail sector (Dabholkar, Thorpe, & Rentz, 1996). Delivery of high service quality has been a target strategy for businesses across sectors in creating a competitive advantage. Perceptions of poor service quality can result in lost customer confidence, or even lost customers, where negative customer experiences can be extremely costly to businesses (Moore & Kelly, 1996). Businesses that are able to demonstrate a high service quality become more competitive in the marketplace by attracting and retaining customers (Berry, 1986; Reichheld & Sasser, 1990). The more service providers are able to evaluate consumer satisfaction and perceptions of service quality, the more they can enhance their attractiveness to other potential customers.

The relationship between perceptions of service quality and consumer behavior has been documented in the literature and has been garnering the attention of businesses in all sectors of service delivery. The 1990s brought on a new paradigm of quality management, where the focus is on delivering services of high quality from a management perspective (Moore & Kelly, 1996). This was the intention for businesses to retain customers in the most competitive industries. Quality management techniques have been a consistent theme in the research in business (e.g., Dobyns & Crawford-Mason, 1991; Walton, 1990) and have more recently extended into the social service organizations (Moore & Kelly, 1996). The increasing attention paid to service quality across sectors suggests service providers must address service quality as a means of acquiring and retaining customers.

There are, however, many problems associated with businesses adopting the orientation to high service quality (Swiss, 1992). Among the concerns is how the quality
is evaluated by the consumer, through difficulty understanding the customer’s point of view when evaluating service quality (Moore & Kelly, 1996). Assessing consumer perceptions of service quality is becoming increasingly important for firms in all sectors as the competition for business increases. That is difficult to do, as these are not directly observable and subjective in nature (Parasuraman, Zeithaml, & Berry, 1985). The intangible definition of service quality makes it difficult to directly assess and collect data with validity and reliability, while challenges arise in using the data in the decision-making process (Boulding, Kalra, Staelin, & Zeithaml, 1993). If reliable and valid data could be collected, such data could be used in evaluating the evolution of service quality in a specific setting or to gauge employee behaviors as a reference for service recovery or advancement (Swiss, 1992). While there is an increasing focus to attend to service quality from customers’ points of view, there is difficulty on behalf of the service delivery firm in evaluating consumer perceptions and using that feedback to make decisions.

Due to the nature of service delivery being a performance and not an object, most services are intangible and distinctions in quality are difficult to judge (Parasuraman, Zeithaml, & Berry, 1985). Precisely how service delivery is perceived, as well as consumers’ evaluations of service quality, makes it difficult to understand and measure due to the inherent intangibility and subjectivity (Zeithaml, 1981). The multi-faceted nature of service quality has long made it difficult to conceptualize, let alone measure. Many researchers have attempted to define service quality as perceived by the service provider and the customer, but have struggled due to the subjective interpretation of the service experience (Brady & Cronin, 2001; Zeithaml, Berry, & Parasuraman, 1988).
Parasuraman, Zeithaml, and Berry (1985) make the statement that “Quality is an elusive and indistinct construct” (p. 41), while others note how it is not easily articulated by either the recipients or consumers (Takeuchi & Quelch, 1983). It is the inherent subjectivity and intangibility that lead many researchers away from defining and measuring service quality.

Existing Measures of Service Quality in General Settings

Many researchers have made attempts to objectify service quality and generate instruments to evaluate the consumer satisfaction with the service experience. A prevalent issue among service quality measures lies in a lack of a common sound conceptual and operational basis on the construct of service quality, a concern that has existed since the beginning of service quality instrument development (Jacoby, 1978). Though there have been many previous attempts at conceptualizing and defining the quality of a service experience, there has been little research exploring actual behaviors of personnel related to quality service delivery until the early 1990s (Boulding, Kalra, Staelin, & Zeithaml, 1993; Ladhari, 2008; Winsted, 2000). Measures of service quality are intended to represent behavioral components of effective and quality service encounters, where the researchers identify behaviors that lead to satisfactory (or unsatisfactory) service encounters (Bitner, Booms, & Tetreault, 1990; Malholtra, Ugaldo, Agarwal, & Baalbaki, 1994). In most cases, those interested in assessing service quality have drawn from a narrow scope of employee behaviors evident in high quality service encounters to generate foundational factors or dimensions of service quality (Bitner, Booms, & Tetreault, 1990; Parasuraman, Zeithaml, & Berry, 1985; Winsted, 2000). The
use of common service quality factors serves as the framework from which service behaviors are most often identified.

One of the most widely used and researched measures of service quality is the SERVQUAL instrument (Parasuraman, Zeithaml, & Berry, 1988). This was one of the first attempts to generate an instrument assessing quality across diverse service sectors. The developers of SERVQUAL conceptually and operationally defined service quality and generated items based on component factors of the construct, resulting in a 22-item scale with demonstrated validity and reliability (Parasuraman, Zeithaml, & Berry, 1988). The SERVQUAL instrument provided the first basic framework for evaluating consumers’ perceptions of the service delivery encounter, beginning with an attempt to operationalize service quality.

The SERVQUAL instrument was developed to be universally applicable to all sectors, covering service and retail organizations (Parasuraman, Zeithaml, & Berry, 1988). Through their research, Parasuraman, Zeithaml, and Berry (1985, 1988) were able to reduce the ten original dimensions of service quality in SERVQUAL to five: 1) tangibles, 2) reliability to perform service accurately, 3) responsiveness in assisting customers, 4) empathy, or individual care and attention, and 5) assurance by the employee to inspire trust and confidence. These five dimensions were concluded to display applicability across sectors reflective of strong service quality. Later attempts were made, using SERVQUAL as the standard, to specify factors or dimensions specific to one sector.
The psychometric properties of the SERVQUAL instrument provide some of the reasons it is widely used and adapted in other settings. The initial development of the SERVQUAL instrument suggested strong total-scale reliability and reliability of the factor structure across several samples (Parasuraman, Zeithaml, & Berry, 1988). The initial psychometric properties of factor structure and internal consistency were evaluated alongside construct validity. Total-scale reliability ranged between .85 and .90 and a factor analysis upheld reliability of a five-factor structure across four subject groups (Parasuraman, Zeithaml, & Berry, 1988). The scale was validated using a combination of conceptual and empirical criteria. Conceptually, face validity and content validity were evaluated by examining the items while empirical testing compared scale scores with an item representing overall quality (Parasuraman, Zeithaml, & Berry, 1988). Overall, the researchers indicate the scale displayed sound psychometric properties and a theoretical foundation to which other instruments should compare.

The introduction of the SERVQUAL measure to the existing service quality literature provided a sufficient attempt at operationalizing service quality into factors that were thought to apply to a range of sectors where service quality is important. In developing the Retail Service Quality Scale, Dabholkar, Thorpe, and Rentz (1996) adopted the SERVQUAL factor structure to reflect the retail sector, where there has been little discussion on a theory-based approach to service quality. Several other researchers (e.g., Akbaba, 2006; Caro & Garcia, 2007; Saleh & Ryan, 1991) have attempted similar instruments in sectors of hospitality, medical, and transportation industries, respectively. The sector-specific scales were developed based on the perceived insufficiency and
shortcomings of the generic SERVQUAL instrument to capture perceptions of quality adequately in different contexts, where service quality is defined by different behaviors (Babakus & Boller, 1992). Thus, the SERVQUAL instrument served as the impetus for the development of other similar, industry-specific measures of service quality.

The most widely recognized contributions of the SERVQUAL instrument were to operationalize service quality and supply a basis for representative factors and corresponding, objective behaviors. These behaviors—central to effective service quality—are termed pro-social organizational behaviors or service behaviors (Bettencourt & Brown, 1997). These service behaviors are direct extensions of the proposed factors that are connected to high service quality, including: 1) tangibles, 2) reliability to perform service accurately, 3) responsiveness in assisting customers, 4) empathy, or individual care and attention, and 5) assurance by the employee to inspire trust and confidence, from Parasuraman, Zeithaml, and Berry’s (1988) research. More specifically, service behaviors are those that a) directly involve employee interactions between customers and co-workers, b) extend beyond role requirements that benefit the organization, and c) involve cooperation (Brief & Motowidlo, 1986; Organ, 1988). The provision of behavioral indicators of service quality led to future research on behavioral change as a means of enhancing service quality (Rice, Austin, & Gravina, 2009). The unique contribution of the SERVQUAL measure centered on the components which constitute service delivery, while many other researchers made attempts to identify service behaviors applicable to a specific industry.
Shortcomings of the SERVQUAL instrument include questions regarding validity, factor structure, and existing, competing theories on the value of customer expectations (Ladhari, 2008). Each domain may have specific factors or dimensions that represent strong service quality, which are questionably applied when trying to examine quality across many domains (Akbaba, 2006). Namely, the factor structure was found to be inconsistent when items were factor analyzed after being collected from multiple sectors (Babakus & Boller, 1992). Other researchers contend the dimensions identified are not appropriate to all sectors, questioning validity of the scale to applications within other domains (Dabholkar, Thorpe, & Rentz, 1996). Instead, several adaptations have followed, suggesting each industry is expected to constitute its own representation of high quality service, to which behaviors should align (Ladhari, 2008). The psychometric properties presented in the development of SERVQUAL have been tested in other domains that have precipitated questions regarding the validity of the factors and applicability to other sectors.

There have been many research calls since SERVQUAL to develop a measure that better defines the meaning of service quality through service dimensions. Several other instruments have been developed in an attempt to fill this void, using the SERVQUAL example and adapting the items to best apply to the focal sector (e.g., Bowers, Swan, & Koehler, 1994; Caro & Garcia, 2007; Vandamme & Leunis, 1992). SERVQUAL remains the standard based on the existing evidence of its sound psychometric properties of reliability and validity (Ladhari, 2008). The measurement framework was adapted several years later in an attempt to validate the five dimensions
(i.e., tangibles, reliability, responsiveness, empathy, and assurance) and advance the applicability of service behaviors across sectors (Parasuraman, Zeithaml, & Berry, 1994). The initial development of the SERVQUAL instrument served to define five critical dimensions of service quality, but few researchers have limited their investigation of service behavior to only these five dimensions.

Other research has identified different dimensions for classifying and assessing service behaviors that might be more applicable than the five presented in SERVQUAL, which may make the generic SERVQUAL instrument not equally applicable across multiple sectors (Winsted, 2000). While the proposed five SERVQUAL dimensions of quality may be generally considered appropriate for multiple settings, importance of each varies based on the service interaction that differs across sectors (Vandamme & Leunis, 1992). The corresponding service dimensions may not be useful in all instances or applicable to all settings based on the relative importance of behaviors as indicators of service quality. Other researchers have identified factors not accounted for by the SERVQUAL instrument, such as perceived control or courtesy which may be relevant to rating service quality in other industries (Winsted, 2000). Ultimately, services may be context-specific and the setting may influence the dimensions of quality that most affect consumer satisfaction (Mowen, Licata, & McPhail, 1993). Each sector defines service quality differently, developing different dimensions that represent strong, positive service behaviors.

The development of service quality instruments in all sectors follows a logical progression beginning with defining the construct (i.e., service quality) and identifying
component dimensions, from which a large pool of behavioral indicators are selected (e.g., Winsted, 2000). The component dimensions serve as a theoretical framework in which quality is operationalized (Sower, Duffy, Kilbourne, Kohers, & Jones, 2001). The dimensions specify behaviors that are included or excluded within the instrument, as behaviors either agree with and fit the dimensions or do not agree and are excluded. In the service quality literature, dimensions serve as the higher-order units of analysis and allow for a comparison across sectors. The use of dimensions instead of a collection of separate behaviors allows for the selection of multiple behavioral indicators, representing each dimension, that are shared across fields where customers are asked to rate service quality (Winsted, 2000). There have been no documented attempts to start with directly observed, specific behaviors instead of higher-order dimensions in the development of a service quality instrument.

Ultimately, there are several perceived benefits to starting instrument development with target behaviors or behavioral indicators instead of factors or dimensions or high quality service. Initially, behavioral indicators serve as operationalized definitions of the dimensions which are already in observable, measurable terms (Sower, Duffy, Kilbourne, Kohers, & Jones, 2001). Secondly, most instrument development exercises begin with qualitative data collection such as focus groups or interviews to define the construct and outline the dimensions (e.g., Churchill, 1979; Sower, Duffy, Kilbourne, Kohers, & Jones, 2001). If the focus groups are experienced in service delivery or receipt, they are already aware of behaviors that are reflective of high quality service and are able to generate items based on these
experiences. Lastly, there is an inherent inter-relatedness among behaviors and dimensions that define quality service experiences (Caro & Garcia, 2007). Beginning with pre-existing hypotheses as to the relationships of items with dimensions may pose concerns in the analysis phase, where assumptions about the relationships between items and dimensions are made that may not be represented accurately in the data (Sureshchandar, Rajendran, & Anantharaman, 2002). Beginning with behavioral indicators of service quality will allow for a later investigation of the underlying dimensions and the inter-relationships among them using factor analytic techniques as opposed to generating items based on presupposed dimensions.

Frequently, identification of behaviors representing good service quality often results in the generation of new service behavior dimensions. Dimensions not identified in SERVQUAL were found within other scales, including but not limited to factors of authenticity, courtesy, formality, and perceived control (Winsted, 2000). Not all instruments, generic or sector-specific, include the same dimensions, deeming it essential for those dimensions relevant to the target sector of focus to be included in assessments. The appropriateness of dimensions and corresponding service behaviors may still be subject to the individual sector to which it is applied, as consumer evaluation of service quality is different across each sector (Vandamme & Leunis, 1992). Service behaviors remain the observable, tangible representation of service quality, which are often present within unique dimensions that represent high quality customer service and customer satisfaction.
Each industry is expected to constitute its own dimensions of high quality service, to which behaviors should align (Ladhari, 2008). Service quality perceptions differ across settings, and that fact makes generic instruments difficult to evaluate the intricacies of service behaviors inherent in specific sectors (Winsted, 2000). A critique of palpable products may not reflect quality in the healthcare setting, where there are limited concrete products of the interaction, yet quality is a particular concern (Vandamme & Leunis, 1992). Not all service behaviors or service quality dimensions are essential in all settings of service delivery, and depending on the sector, consumers may place more importance on some aspects of service quality (Ladhari, 2008). Each industry-specific measurement of service quality should consist of behaviors that are tangible representations of service quality as they apply to that setting (Akbaba, 2006). Overall, the setting serves to dictate the behaviors that are essential in quality service delivery, which makes a single, universal instrument difficult to apply (Babakus & Boller, 1992). An investigation of service quality in a unique setting, such as healthcare, should be rooted in foundational dimensions of high quality service and follow behaviors relevant to the healthcare setting.

**Service Quality in the Healthcare Setting**

Adaptation of items from the SERVQUAL instrument may not provide sufficient assessment of service quality across some service industries, in particular healthcare (Brown, Churchill, & Peter, 1993). The dimensions that serve as the framework for the SERVQUAL instrument (i.e., 1) tangibles, 2) reliability to perform service accurately, 3) responsiveness in assisting customers, 4) empathy, or individual care and attention, and
5) assurance by the employee to inspire trust and confidence; Parasuraman, Zeithaml, & Berry, 1988) may not be appropriate when there are limited tangibles or when the accuracy of the service is unknown to the client (Bitner, Booms, & Tetreault, 1990). The initial foundation for the SERVQUAL instrument may not readily apply to the healthcare setting, while omitting essential service quality factors in need of investigation. There are other factors or dimensions that may be undervalued in the generic instruments, ignoring behaviors that healthcare consumers may value in the service encounter (Bitner, Booms, & Tetreault, 1990). Thus, the general service behavior assessments lack specificity in targeting important behaviors representing service quality in some settings. Together, the approach of using a universal assessment of service behaviors may not be appropriate. Instead, an industry-specific approach with behaviors tailored to the service encounter may be preferable.

Healthcare is a good example of an industry in which the generic service behavior assessments may not readily apply. Health services organizations are meaningfully different from other sectors on the basis of design, mission, and the amount of resources that flow to and from them, which may be attributable to differing service encounter expectations (Scotti, Harmon, & Beson, 2007). The medical setting differs from the other sectors in terms of service behaviors as the service delivery comes from qualified professionals, not unskilled workers or general laborers (Hill & Motes, 1995). Doctors generally have long-term, formal relationships with patients and there is relative complexity in communication patterns and customer problems (Bitner, Booms, & Tetreault, 1990). Thus, healthcare organizations present meaningful differences from
other service organizations in terms of measurable behaviors representing service quality (Scotti, Harmon, & Behson, 2007). Few generic instruments are able to be applied effectively in the healthcare setting, calling the validity of the existing instruments into question.

In the healthcare setting, the identification of behaviors that are representative of service quality is particularly problematic. The challenge of assessing service quality and identifying service behaviors in healthcare is based largely on the notion that consumer expectations and satisfaction with healthcare are difficult to pinpoint (Kobayashi, Takemura, & Kanda, 2010). For a long time, it was not the expectation for hospitals to excel in the domain of service quality, leading some institutions to have difficulties in assessing the service quality for patients (Mekoth, George, Dalvi, Rajanala, & Nizomadinov, 2012). In the past, evaluations of service quality in healthcare have focused on clinical effectiveness and patient safety (Doyle, Lennox, & Bell, 2013). These have long been the metrics used in judging healthcare settings.

It has been through the increasing competition in the healthcare marketplace that special attention has been paid to patient experience and perceived employee service behaviors (Scotti, Harmon, & Beson, 2007). Similar to consumer experiences in retail and hospitality, satisfied patients are expected to become loyal customers. The inclusion of patient experience to other indicators of quality in healthcare—clinical effectiveness and patient safety—is justifiable in itself (Doyle, Lennox, & Bell, 2013). Quality evaluations in healthcare have most often focused on service personnel such as nurses and therapists (i.e., allied health staff) as distinct from physicians (Kennedy, Caselli, &
Berry, 2011) based on the complexity and nature of patient-physician problems (Bitner, Booms, & Tetreault, 1990). Healthcare organizations are becoming increasingly interested in assessing patient satisfaction and perceptions of service quality as a way to maintain a competitive edge over other healthcare organizations.

The changes in healthcare reform legislation have also placed increasing focus on quality and value of what healthcare organizations represent to consumers (Kennedy, Caselli, & Berry, 2011). Patient-centered care has long been a critical component of some healthcare organizations’ offerings, where the ultimate outcome of healthcare has consistently focused on patient satisfaction (American Nurses Association, 1995; Donabedian, 1966). Based on current trends, hospitals and other organizations are becoming accountable for patient experiences, while maintaining clinical effectiveness at the same time (Kennedy, Caselli, & Berry, 2011). Research indicates the enhancement of service quality and consumer satisfaction contributes to cost efficiency, above and beyond inflating costs (Scotti, Harmon, & Beson, 2007). As the need for providing quality services increases, so does the need for proactive evaluation of those services that are specific to the healthcare industry.

An early attempt to develop a healthcare-specific instrument was made by Vandamme and Leunis (1992), shortly after the development of SERVQUAL (Parasuraman, Zeithaml, & Berry, 1988) and as a specific response to its presented findings. The instrument developed by Vandamme and Leunis (1992) utilized the five dimensions from SERVQUAL (i.e., tangibles, reliability to perform service accurately, responsiveness in assisting customers, empathy, and assurance) but adjusted the items to
reflect hospital service quality. The resulting instrument assessed the SERVQUAL dimensions through items representative of the healthcare sector in terms of both physicians and allied staff together as service providers. A review of the psychometric properties yielded a factor structure that differed from the SERVQUAL instrument and moderate reliability (i.e., $\alpha = .58-.75$ for each dimension assessed), leaving questions about the validity of the scale with a different sample (Vandamme & Leunis, 1992). The researchers provided a restricted discussion regarding the psychometric properties of the scale, which was limited to internal consistency reliability, content validity, and construct validity. Because evidence of validity was not strong in this setting, the researchers concluded that SERVQUAL’s dimensions were unable to capture all of the relevant service quality dimensions from the healthcare sector (Vandamme & Leunis, 1992).

A later attempt to address the need to evaluate service quality in the healthcare setting resulted in the Key Quality Characteristics Assessment for Hospitals (KQCAH), originally developed to identify and operationalize dimensions of patient satisfaction (Sower, Duffy, Kilbourne, Kohers, & Jones, 2001). The KQCAH evaluates the multiple dimensions of hospital service quality tapping only on patient perceptions as indicators (Ladhari, 2008). The KQCAH was able to display high levels of content and construct validity through qualitative examination by the researchers, as well as high reliability (i.e., $\alpha > .70$ for each dimension assessed) when its psychometric properties were evaluated (Sower, Duffy, Kilbourne, Kohers, & Jones, 2001). The development of the KQCAH cultivated the interest in operationalizing service quality in the healthcare setting utilizing 8 dimensions derived from 75 items: 1) respect and caring, 2)
effectiveness and continuity, 3) appropriateness, 4) information, 5) efficiency, 6) effectiveness-meals, 7) first impression, and 8) staff diversity (Sower, Duffy, Kilbourne, Kohers, & Jones, 2001). On the KQCAH, respondents are asked to report service quality in terms of overall staff, without accounting for professional role or the nature of the interaction. These dimensions of service quality represented in the KQCAH, overlap with other instruments developed specifically for application in the health services sector (e.g., Caro & Garcia, 2007; Vandamme & Leunis, 1992), suggesting agreement and emphasis on these behavioral indicators as they reflect high quality service in healthcare.

The most recent attempt to measure perceived service quality in a healthcare domain came from Caro and Garcia (2007). Their instrument evaluated service quality in urgent transport service in Europe, which may have questionable applications to other healthcare disciplines but contributed to the dimension identification and analysis of service quality in healthcare settings in the United States. Caro and Garcia (2007) adapted items based on previous scales like SERVQUAL (Parasuraman, Zeithaml, & Berry, 1988) and the Retail Service Quality Scale (Dabholkar, Thorpe, & Rentz, 1996), to best apply to the healthcare sector. The four dimensions that were evaluated were those related to 1) personal interaction (i.e., conduct, expertise, and problem-solving), 2) design (i.e., range of services provided and operating time), 3) the physical environment (i.e., tangibles and information shared), and 4) outcome (i.e., punctuality and belief in a good service outcome), all of which were strong indicators of service quality (Caro & Garcia, 2007). The novel contribution of this instrument goes beyond the dimensions evaluated, into employing an advanced analysis of inter-relationships among the behaviors and
dimensions. The behavioral indicators present in the items were reflected in the dimensions, which demonstrated a statistical relationship with service quality based on confirmatory factor analysis techniques (Caro & Garcia, 2007). This was the first documented investigation of instrument psychometric properties beyond simple reliability and validity, providing a standard to which future instruments should be compared. Overall, the instrument provided a unique contribution to the service quality literature; however, application of the scale and its dimensions remains questionable to hospital settings in the United States (Caro & Garcia, 2007, Ladhari, 2008).

The existing scales and accompanying evidence suggest the evaluation of psychometric properties beyond the data provided regarding reliability, validity and factor structure is limited; namely missing an item-level analysis using Item Response Theory (IRT) principles. The instrument developed by Vandamme and Leunis (1992) was the first of its kind, generated to evaluate service quality in the healthcare setting using the SERVQUAL factor structure, while the KQCAH was the first to demonstrate adequate scale properties of validity and reliability (Sower, Duffy, Kilbourne, Kohers, & Jones, 2001). Caro and Garcia (2007) made the most recent developments employing confirmatory factor analytic techniques to an instrument evaluating urgent transport service in Europe, but its application to hospitals in the United States remains questionable. Together, the existing measures of service quality have demonstrated adequate reliability, validity, and factor structure, but no evidence suggests any have gone on to more advanced, IRT evaluations of how service quality is captured within the instruments while confirming the underlying factor structures.
Each of the existing service quality instruments designed for the healthcare setting is based on patient report of the service encounter and evaluate staff as a whole, not accounting for differences in professional roles or interactions. Collectively, attention to service quality has focused on consumers’ perspectives, mostly due to the nature of service quality being inherently subjective (Zeithaml, 1981). This is also indicative of the role played by consumers and their level of involvement in the service delivery process, particularly within the healthcare setting (Mekoth, Georgy, Dalvi, Rajanala, & Nizomadinov, 2012). Patients as consumers tend to be highly involved, though they often know little about accuracy of services and there are far less concrete results by which to identify service quality in healthcare (Bitner, Booms, & Tetreault, 1990). In their review of methods of measuring patient satisfaction in healthcare, Ford, Bach, and Fottler (1997) included qualitative and quantitative methods, including patient surveys, but neglected to identify the potential for employee self-report. The existing measures continue to rely on consumer perceptions of service quality, where the nature of the interaction and consumption of service make it extremely difficult for the consumer to be an accurate judge of service quality (Zeithaml, 1981). The continued use of consumers to evaluate the service encounter, given their difficulty to accurately judge quality, leads to questions as to how better assess behaviors indicative of high quality service.

Previous measures have called upon the consumer to provide an overall evaluation of the service encounter as a means for judging quality. The existing consumer-report instruments are most helpful in evaluating a single patient’s complete service encounter, but may not be indicative of individual service providers’ behavior as
a quality indicator in healthcare (Vandamme & Leunis, 1992). In other avenues of research, there is a distinct difference between satisfaction with respect to a specific transaction and a customer’s global evaluation of a service (Holbrook & Corfman, 1985; Olshavsky, 1985). Thus, overall customer satisfaction ratings may apply to their summative evaluation of the service encounter instead of formative interactions between any number of service providers. Particularly as they apply to the healthcare setting, differences in service quality attributable to an individual person is not possible using a global rating of consumer satisfaction (Holbrook & Corfman, 1985). The overall impression of service quality can be shared between many individuals as patients encounter several trained, qualified professionals delivering various services due to the compartmentalized structure of the healthcare setting (Hill & Motes, 1995). Together, an overall rating of service quality is intended to reflect the quality as a whole and is not attributable to separate personnel involved in service delivery.

It is the possible quality differences across persons performing the service delivery that make it difficult for institutions to identify weaknesses in individual service behaviors when using a single consumer-completed service quality measure. The inseparability of global service quality ratings to individuals or the interrelatedness of service delivery in healthcare that relies on multiple interactions may make such an overall evaluation of service quality not applicable to individual employees, especially given the labor-intensive organizational framework of healthcare settings (Vandamme & Leunis, 1992). The global evaluations provided by patients can be impacted by any number of individuals within the organizational framework, providing a service quality
rating that may not be accurate (Olshavsky, 1985). Extrapolating an evaluation of all staff to the differences in interactions between patients and physicians and allied staff is inappropriate, given the complexity of patient-physician interactions (Bitner, Booms, & Tetreault, 1990). From the healthcare standpoint, patients serve in the role of consumers, where accuracy of consumer report may not be as accurate as expected (Zeithaml, 1981). An inaccurate or mis-representative evaluation of service quality can have striking impacts on employees and healthcare systems, which may lead healthcare providers to identify and address individual differences or shortcomings in such behaviors.

For healthcare institutions to gauge growth in service quality the use of a single, global evaluation of employee behavior provided by patients may not be the most appropriate; instead, an employee-based assessment may provide a more sensitive insight into behaviors relating to high service quality. The implementation of employee self-report measures can be useful in gauging employee perceptions of their own behaviors and reflective of employee values on culture (Steers & Porter, 1991). There has been no evidence of an attempt to evaluate the quality of a healthcare service encounter from an employee’s perspective, or to provide a way to gauge employees’ self-perception of their own behavior throughout the service encounter. The assessment of employee self-perceptions of service quality has been found to be more direct and suitable in examining organizational behavior compared with customer report (Steers & Porter, 1991). Further, the voices of employees are particularly relevant when evaluating service quality (Schneider, Parkington, & Buxton, 1980), but are most effective in identifying practices and procedures of high quality service (Yavas, 2007). Employees and their customers
perceive quality of service similarly, and employee perceptions of organizational service-related practices and procedures are strongly related to customer perceptions of service quality (Schneider, Wheeler, & Cox, 1992). Together, an instrument that evaluates employee self-perceptions may provide insight into employee behavior and growth in high quality service behaviors.

There remains a need for an instrument to address service quality within the scope of behaviors critical to the healthcare setting, provided by employees, while demonstrating adequate psychometric properties. Of the few attempted measures of service quality in the healthcare setting, all of which follow a similar instrument development and validation process (e.g., Caro & Garcia, 2007; Sower, Duffy, Kilbourne, Kohers, & Jones, 2001; Vandamme & Leunis, 1992). The instrument development process described in the service quality literature provides a sound framework for item development and analysis to ensure minimal validity and reliability. To generate a measure that stands up to the strictest criteria for sound psychometric properties, instrument development frameworks will be drawn upon from both the existing service quality literature as well as psychology and sociology, where many instrument development manuscripts are published.

**The Instrument Development and Refinement Process**

The instrument development process varies widely dependent upon context and purpose, but often follows a common process. The instrument development procedure proposed by Churchill (1979), which stresses the collection of data to uphold validity and
reliability assumptions, is mirrored in other measure development research such as Sower and colleagues’ (2001) KQCAH scale and Dagger, Sweeney, and Johnson’s (2007) development of another service quality instrument. These resources provide the precedence and framework from which other measure developments should follow.

When developing an instrument, the initial phase is to specify the domain of the construct (Churchill, 1979), which is very challenging with so many dimensions of service quality from which to choose (Akter, D’Ambra, & Ray, 2013; Sower, Duffy, Kilbourne, Kohers, & Jones, 2001). This procedure involves defining what the measure will assess and setting limits for what is included and what is excluded, which has been established well documented in the healthcare setting (Dagger, Sweeney, & Johnson, 2007). The previous literature provides an adequate description of the domain of service quality and a framework for its assessment in the healthcare setting.

Flowing from the definition of the construct and the criteria for inclusion, the next step is to operationalize the construct and form representative behaviors of the dimension of interest (Churchill, 1979). With the dimension of interest being service quality in the healthcare setting, item development should follow the conceptualization of service quality in hospitals (Akter, D’Ambra, & Ray, 2013; Sower, Duffy, Kilbourne, Kohers, & Jones, 2001). Items should be developed using the specific underlying factor structure as a framework. The conceptualization of service behaviors could be based on focus group data collection, which is a common exploratory method to postulate a framework for instrument development when a purposive sample is selected (Dagger, Sweeney, & Johnson, 2007; Morgan, 1997). Once the construct is designed and the conceptual
Item Development and Pilot Testing

Once the construct is defined and operationalized, items should be developed so they fit within the construct and capture the specified domain accurately (Churchill, 1979). In the current case, the target institution has identified essential service behaviors instead of pre-determined dimensions of service quality, so item development is based on the expected service behaviors. The item pool should initially cover an array of the construct since the original items will be refined leading to the final measure (Churchill, 1979). It is common to be inclusive in the item generation process, while selectivity can follow once data become available as a means of producing the final instrument.

Therefore, the initial item pool was considered all-inclusive and included an array of behaviors relative to the construct under investigation.

Once the item pool is developed, pilot testing ensues (Dagger, Sweeney, & Johnson, 2007). The process of pilot testing and analysis following item development is to ensure both content and construct validity (Akter, D’Ambra, & Ray, 2013). The items must represent the construct under investigation (i.e., service quality), while assessing a representative sample of the behavior domain. Pilot testing results are often analyzed to assess instrument reliability and item factor structure to be confirmed in full-scale administration (Akter, D’Ambra, & Ray, 2013). The instrument development process is formative, as each administration and testing must yield data that provide evidence for change and refinement (Churchill, 1979).
Instrument Refinement and Confirmatory Analyses

When the pilot testing is completed, the results should be used formatively to adjust the instrument as necessary to enhance psychometric properties and purify the measure (Churchill, 1979). This process is typically completed by collection of new data and confirmation of the pilot results (Akter, D’Ambra, & Ray, 2013). If any changes to the measure are made, new data should be collected and analyzed in a way identical to the pilot to document the formative change of the instrument. It is common that some items are dropped or added between pilot and full scale testing sessions based on poor item performance or other problematic issues that arise, leading to the instrument to undergo several phases prior to finalization (Heinz, Kassel, & Smith, 2009). Regardless, a larger sample involved in the full scale administration will allow the researcher to confirm proposed factor structure and item performance suggested by the pilot administration (Akter, D’Ambra, & Ray, 2013). The confirmatory analyses seek to establish, on a much larger scale, the psychometric characteristics of the instrument once it has been refined.

Methods and Models Employed in the Psychometric Property Examination

The instrument development processes prescribed by Churchill (1979) and Dagger and colleagues (2007) diverge in the method to evaluating the instrument’s psychometric properties. The analyses employed to validate the measure may include a combination of descriptive statistics and reliability coefficients, exploratory and confirmatory factor analyses, and an Item Response Theory based Rasch analysis. These
analyses are also able to represent how the service quality construct is captured by the measure.

**Item Analysis**

There are several common methods for item analysis to address concerns with validity and reliability in the instrument development process. Many of reliability analyses involve Classical Test Theory (CTT) approaches utilizing the sample form which the data is drawn (Crocker & Algina, 2008). Chief among the item-level analyses are checks for validity and reliability that include an evaluation of internal and external validity, as well as internal consistency or coefficient alpha (Cortina, 1993). Internal validity is achieved when the instrument measures what it intends to measure, while external validity is representative of the generalizability of the results (Cronbach, 1971). Among the concepts explored in internal validity are content and construct validation.

**Validity.** Content validity refers to the extent to which the items on an instrument represent a domain or construct of specific interest (Allen & Yen, 1979). Face validity is applied to determine if the items appear to resemble the construct of interest (Churchill, 1979). The items must not only look like they tap into the domain of focus, but they must be representative of the domain as well. Along with content validation, there should also be construct validation for a new instrument. Construct validation occurs when the newly-developed items provide scores that correspond with existing measures of the same focus (Wright & Stone, 1979). For an instrument to have met the criteria to be valid, it must demonstrate these basic tenets of internal validity.
External validity can be enhanced through sample diversity to increase the generalizability of the results (Kazdin, 2003). Since external validity applies to how well the results can be generalized outside the research on the sample, there are limitations on this when applying an instrument to one sector (Crocker & Algina, 2008). The limits of generalizability of the results for an instrument developed specifically for the healthcare setting are expected to be only applicable to other healthcare settings. The interest of developing an instrument with external validity lies in having results of an instrument administered in one healthcare setting may be generalizable to other settings, such as hospitals or outpatient clinics, if the dimensions of service quality are equally appropriate.

**Reliability.** Reliability refers to the consistency or reproducibility of the results from a testing administration (Allen & Yen, 1979). Coefficient alpha is one of the most often used indices of internal consistency and represents the average of all possible split-half coefficients that would be obtained if the instrument was divided into all possible combinations of dividing the test in half (Cronbach, 1951). Coefficient alpha is the most basic reliability statistic and is helpful in identifying problematic items when examining the coefficient with item removed (Churchill, 1979). The coefficient is presented on a scale ranging from 0 to 1, where scores closer to 1 indicate strong internal consistency and values less than .70 are typically inadequate and unreliable measures (Cronbach, 1951). Reliability analyses were conducted because the item statistics regarding reliability provide insight into the replication of scores, as consistency is sought in measures with strong psychometric properties.
Exploratory Factor Analysis

The purpose behind Exploratory Factor Analysis (EFA) is primarily exploration of the data to help simplify results into theories, which lead to the classification of EFA as typically theory-generating (Tabachnick & Fidell, 2001). Factor analysis can be used to explore patterns in data, confirm hypotheses, and reduce many variables into a manageable number that might have value in interpretation (Norman & Streiner, 2003). EFA is frequently used in the instrument development process to assist in assessing dimensionality and identifying constructs within items (Costello & Osborne, 2005). In the current case, the implementation of EFA techniques will be used to determine the number of underlying factors in the target service behaviors and to establish existing relationships between the items.

In EFA, the researcher has no specific expectations regarding the number or nature of underlying constructs or factors (Thompson, 2004). The current service quality and behavior literature suggests a different number of constructs, ranging across studies from one to ten or more (Winsted, 2000). EFA will be applied to the pilot testing data to evaluate the number of constructs evident in the items. Since the item development will take place using the institution’s service behaviors as a guide, there is no conceptual framework or existing underlying factor structure in place. The EFA is anticipated to identify any foundational structure underlying the items.

The process of EFA was first proposed by Spearman (1904) and has gained in use across the social and healthcare sciences over the past 15 years (Pett, Lackey, & Sullivan, 2003). In EFA, the variance within a given variable (i.e., item) can be explained by the
variance that is unique to that individual variable, as well as a smaller number of underlying, common factors (Thompson, 2004). A factor represents a linear combination of the observed variables and indicates the presence of a dimension, distinct from other factors in the solution (Tabachnick & Fidell, 2001). The underlying factors serve to identify and explain inter-relationships among variables (Kim & Mueller, 1978), while the factors maintain common characteristics (Nunnally & Bernstein, 1994). The analysis utilizes the correlations or covariances among items and identifies factors sequentially, removing those first that explain the most variation within items (Thompson, 2004). The purpose of using EFA in the current investigation is to identify the amount of variance in the items that can be explained by removing common factors. Identification of common factors within the items may assist in detecting multiple dimensions and comparing with other factors prominent in the service behavior literature.

Some of the major critiques of EFA lie in its frequent misuse and disregard for limitations. The incorporation of factor analysis into statistical programs has made it readily available to people who are unfamiliar with the methods (Kline, 1994). Failure to acknowledge the limitations of the technique lead it to be one of the more widely used and misused statistical techniques (Norman & Steiner, 2003). There are no explicitly-stated directions for EFA, given the multitude of choices and options available to a researcher (Costello & Osborne, 2005). There are many decisions that must take place when employing EFA, from estimation method and rotation to the number of factors extracted. The current investigation and use of EFA techniques will follow some common best practices using the data that is obtained during pilot testing, utilizing the
appropriate estimation method and rotation to generate the most accurate and reliable results.

One major concern is that EFA results are often misinterpreted because of failure to reach an appropriately-sized sample. Costello and Osborne (2005) suggest that though no strict rules regarding sample size exist, factors can be accurate, reliable representations of the data if the sample size exceeds a 10:1 ratio of respondents to items. The pilot testing phase is expected to exceed the necessary sample size for reliable results of the EFA. Failure to reach an appropriate sample size will be noted and results will be interpreted with caution.

**Confirmatory Factor Analysis**

Contrary to EFA, Confirmatory Factor Analysis (CFA) is traditionally theory-testing in which there are expectations set by the researcher who sets out to determine if the initial theory is upheld (Thompson, 2004). CFA assesses the extent to which the hypothesized structure of identified factors conforms to fit the data (Nunnally & Bernstein, 1994; Pedhazur & Schmelkin, 1991). Thus, the CFA is employed to determine how the proposed factor structure aligns with the data. When coupled with the results of an EFA, a confirmatory analysis tests the dimensions of the construct identified in the exploratory analysis (Pett, Lackey, & Sullivan, 2003). As part of the current investigation, the CFA will be applied to the full-scale administration of the instrument as a means of verifying the factor structure identified in the EFA to further establish validity of the instrument.
CFA seeks to confirm the validity of the hypothesized model by applying statistical analyses determining data fit with the model (Schumacker & Lomax, 2010). The hypothesized model restricts the interrelationships between variables and factors to what was identified as part of the EFA (Dimitrov, 2009). The hypothesized model will be derived from the exploratory analysis conducted on the pilot testing data. Using the results of the EFA will complete the first step of the CFA process, which is model specification (Schumacker & Lomax, 2010). The remaining steps in CFA comprise the evaluation of model adequacy (Dimitrov, 2009), which consist of identification, estimation, testing, and modification (Schumacker & Lomax, 2010). The necessary steps will be taken to follow the appropriate analysis method outlined by Schumacker and Lomax (2010), to verify the factor structure of the data obtained in the full-scale administration.

Adequacy of the model will be addressed by examining goodness-of-fit indices. One of the most common goodness-of-fit indices is the Chi-square, which tests the residual between the covariance matrices for the sample and the model expected by the population (Dimitrov, 2009). The Chi-square, however, does not provide enough evidence alone to indicate satisfactory or unsatisfactory model fit (Bentler & Bonnett, 1980), so other indices will be considered. Among the other common goodness-of-fit indices are the goodness-of-fit index (GFI) and the AGFI, which is the goodness-of-fit index adjusted for population (Dimitrov, 2009). The standardized root mean square residual (SRMR) and the root mean square error of approximation (RMSEA; Steiger, 1990) are also commonly cited statistics used to evaluate the model-data fit (Schumacker
& Lomax, 2010). Together, an array of indices will be used to determine the fit between the full-scale data and the hypothesized model, generated from the preceding EFA. Modifications to the model that suggest stronger model-data fit will be considered only if they can be backed by existing theory.

**Rasch Analysis**

Item Response Theory (IRT) consists of a family of models that use item and person characteristics to predict observed responses of individuals based on their latent traits (de Ayala, 2009). Rasch analysis (Rasch, 1960) is part of the IRT family and was developed to counteract limitations that exist in Classical Test Theory (CTT). The Rasch model is also called the One-Parameter Logistic (1 PL) model, and can be viewed as a special case of the 2 PL (i.e., that includes item difficulty and discrimination) with all items having the same discrimination parameter of 1 (de Ayala, 2009). In the Rasch model, item difficulty and the proportion of examinees correctly responding is a function of the examinees’ ability and the difficulty of the individual item (Hambleton, Swaminathan, & Rogers, 1991). The estimates from the model are represented in logit values, which report the relative differences between ability estimates and item difficulties along an equal interval of measurement (Bond & Fox, 2007). The use of logits allows for item difficulty and person ability to be reported on the same interval measurement scale. The application of IRT or Rasch analysis to existing instruments of service quality has yet to be seen, likely due in part to the complex nature of the analysis or the challenging conceptualization of item difficulty and person ability as it applies to a service quality instrument.
When using ordered, polytomous data in Rasch analysis, the Rating Scale Model (RSM) is the only appropriate model to apply (Andrich, 1978). The RSM requires that the available responses are in ordered categories (e.g., from “Never or Almost Never” to “Always or Almost Always”), and between them are critical points, or thresholds (de Ayala, 2009). These thresholds separate the categories of responses, and researchers are able to examine the characteristics of rating scale functioning using these thresholds. That is, the RSM has person and items estimates, but also thresholds that represent the level at which the likelihood of being in one response category on the scale (i.e., below the threshold) is exceeded by the likelihood of being in the next category (i.e., above the threshold; Bond & Fox, 2007). The use of the RSM is appropriate when each item on the instrument is based on the same response scale (Andrich, 1978), with higher scores indicating more frequent engagement in the service behaviors. The Rasch analysis will be used to evaluate the range of item difficulties, the distribution of person responses, and the structure of the response categories. This research will apply Rasch measurement techniques, specifically the RSM, to examine the psychometric properties of the newly-generated service behavior instrument.

Also, data from the Rasch analysis is anticipated to provide evidence of unidimensionality when the Rasch residuals are factor analyzed using Rasch Factor Analysis (RFA; Linacre, 1998). The assumption of unidimensionality of the instrument is a central tenet of the Rasch model, and failure to meet this assumption provides incorrect estimations and distorted results (Bond & Fox, 2007). The assumption of unidimensionality is not met if items present response patterns that indicate similarity
between some items that differs from others. The RFA is a Principal Components factor analysis on the residual values between the obtained data and the Rasch measurement model (Linacre, 1998). There is expected to be some deviation between the data and the model, so the residuals are analyzed to determine patterns that may represent secondary dimensions within the items that detract from model-data fit.

Not only does the RFA determine if the assumption of unidimensionality was met, it is also helpful in further identifying underlying constructs or components within the items (Linacre, 1998). The results of the RFA will assist in validating any underlying factors, to which the previous CFA should produce similar results. Together, the RFA and CFA seek to establish concomitant factor structures present in the data based on response patterns. Such response patterns could provide evidence to refute the assumption of unidimensionality or suggest underlying constructs or factors present within the instrument. In the event distinct factors are discovered as part of the CFA, items may be grouped together based on factors for separate Rasch analyses to appropriately account for unique dimensions (Linacre, 1998). Applying the Rasch model to subsets of data is most appropriate when each subset statistically and theoretically represent a unique construct (Bond & Fox, 2007). Thus, the CFA and the RFA will be used to determine if the instrument addresses a single factor of service quality in the healthcare setting or if there are constructs within the domain of service quality that are evident within the items.

Finally, the Rasch measurement approach was used to investigate potential response invariance or bias by evaluating Differential Item Functioning (DIF). Simply,
DIF is a form of item bias in which significant differences are found between groups of respondents (Lord, 1980). Item bias, or invariance, is investigated for each of the items by calculating the estimates of groups separately. The Rasch estimates can be compared across two or more distinct groups of interest and tested for statistically significant differences, reflecting the presence of DIF (Bond & Fox, 2007). The presence of DIF suggests that items have different meanings for the different groups, which is evidence of underlying bias within items. Subjecting an instrument to testing for item bias using DIF analysis has yet to be seen in the service quality literature.

In summary, the existing service quality literature supports both the strengths and limitations of previously-developed instruments in both healthcare and other sectors. Strengths and limitations are evident both on a theoretical basis and in terms of psychometric analyses used in validation. Specific to the healthcare setting, existing measures have focused on patient report of the overall service encounter and failed to account for individual employees and their self-perceptions of high quality service behaviors. An existing research framework for the instrument development process has been presented and discussed as it will be applied to the development of a new instrument, followed by multiple analyses aimed at refinement and validation.
CHAPTER III

METHODOLOGY

The presentation of the methodology is divided into two parts: one representing the development of the instrument and pilot testing phase, with the second representing the full-scale administration.

Part I: Pilot Testing Phase

The discussion of the pilot testing phase begins with a description of the process of item development and flows into the presentation of participants, procedure, and the analysis plan based on the pilot data collection.

Participants

Participant data was collected from a large healthcare institution and its affiliated facilities in the Midwestern United States. All data collection efforts, from participant recruitment to instrument administration, were run through the institution’s Office of Human Resources and Office of Patient Experience to avoid third-party issues of selection and administration. Employees classified as physicians or temporary employees were excluded from participation, given contractual rights and relevance of the target behaviors to each population. All remaining employees were identified as subjects for the current investigation through their participation in the one-day new-hire orientation offered by the institution within the study timeline. A single new-hire orientation was utilized to elicit participants for the pilot testing phase, which provided a sample of 570
employees that were distributed an email message containing the electronic link to the pilot instrument within their first six months of service.

Participants were expected to represent a diverse range of demographic characteristics, from gender and race or ethnicity to education and number of years of experience. It is expected that the sample drawn was fairly evenly distributed between genders and racial or ethnic backgrounds, as employees range from practitioners with advanced training and experience to general laborers who require minimal experience or positional qualifications. Levels of education and career experience also differ, where some positions included require undergraduate or graduate training while others may require only a high school education. Some participants have prior experience in similar positions prior to employment in the current setting. Participant ages are expected to range from 21 to 65 years old; however, those outside this range were not excluded. The inclusive nature of the data collection was expected to provide a diverse sample of new employees, representative of various ages, racial or ethnic backgrounds, previous experience and levels of training.

Measure

The generation of the pilot instrument began with item development, which is described in detail in the first section. The details of the pilot instrument, once items are composed, are included in the second section.

**Item development.** The service behaviors that are the focus of the current research were provided by the healthcare setting in which the research took place. The nine behaviors were identified through qualitative research methods, where focus groups
were conducted with members coming from patient advisory councils. The participants included current and previous patients, as well as patients’ family members who had experience in the setting. The healthcare institution has set selection criteria for the patient advisory councils, where participation is limited and selection is by invitation. Participants in the patient advisory councils are confidential and the service behaviors are not attributable to any specific group.

A set of nine target service behaviors were selected because they were behavioral indicators of good service quality identified by the focus groups, while the institution agreed with the focus group feedback. The institution has based their focus on patient-centered, high-quality service using the following nine Expected Service BehaviorsSM.

1. Acknowledge the other person
2. Introduce self and role
3. Use person’s preferred name and greet warmly
4. Clearly communicate expectations
5. Offer to resolve concerns or forward to the right person
6. Use active listening
7. Show empathy
8. Use common courtesy
9. Offer to help

Of the behaviors, the first two deal with the beginning of the caregiver-patient interaction where greeting and introductions take place. Five service behaviors are concerned with caregiver-patient interaction and communication, where the caregiver is
expected to listen and communicate with the patient regarding expectations, concerns, and offering assistance. The remaining two service behaviors are more open-ended, which involve behaviors associated with interpersonal interactions and relationship building. Together, the nine target service behaviors cover the entire caregiver-patient interaction and are applicable to all caregivers, regardless of the level of patient interaction involved in their roles. The institution was seeking to define service quality using behaviors related to patient-centered care, and there was strong correspondence between behaviors that the focus groups and the institution concluded were reflective of service quality, represented in each of the nine target service behaviors.

The target service behaviors have been implemented within the institution to a limited extent. Mainly, the target service behaviors have been incorporated as part of the employee service recovery program. As the institution’s mission has adapted to include delivering high service quality, they have incorporated the service behavior training for all new employees. The nine service behaviors are well-defined by the institution and examples are provided for employees to reference as part of and following the training. Materials and publications presented throughout the institution serve as cues to employees, reminding them of the Expected Service BehaviorsSM and employee impact on patient experience.

The service behaviors were used throughout the item development process to maintain focus and intent of the target behaviors. Each of the nine target service behaviors presented by the institution are specific and explicitly stated so there is minimal subjectivity in determining items representing the behaviors. Due to the nature
of the Expected Service Behaviors<sup>SM</sup>, some were already operationalized (e.g., Introduce self and role, Offer to help), while others needed to be broken down into observable behaviors (e.g., Show empathy, Use active listening). The first stage of item development took place using the service behaviors as the guide.

Eight of the nine service behaviors were represented in two or more items without obvious redundancy. Only one behavior (i.e., Acknowledge the other person) was identified by a single item given the nature of the target behavior. Items with strong face validity in regards to the service behaviors will be monitored so there is minimal overlap, in the event the service behaviors represent similar service qualities. Those Expected Service Behaviors<sup>SM</sup> that were operationally defined by the institution (e.g., Introduce self and role, Use the person’s preferred name and greet warmly, and Offer to help) are represented by three items each, while behaviors with more components (e.g., Show empathy or Use common courtesy) consist of four or more items. A total of 37 items were be generated strictly from the institution’s target service behaviors, and the breakdown is presented in Table 1.

Table 1.  

| Expected Service Behaviors<sup>SM</sup> and Number of Corresponding Items on Pilot Instrument |
|---------------------------------------------------------------|------------------------------------------|
| Behavior                                                      | # of Items on Pilot Instrument |
| Acknowledge the other person                                  | 1                                      |
| Introduce self and role                                       | 2                                      |
| Use person’s preferred name and greet warmly                   | 3                                      |
| Clearly communicate expectations                               | 2                                      |
| Offer to resolve concerns or forward to the right person      | 5                                      |
| Use active listening                                          | 5                                      |
| Show empathy                                                  | 7                                      |
| Use common courtesy                                           | 9                                      |
| Offer to help                                                 | 3                                      |
| Total Pilot Items based on Expected Service Behaviors<sup>SM</sup> | 37                                     |
In addition, nine items were included outside the scope of the expected service behaviors as a means of capturing service quality in the target institution. There are behaviors that have demonstrated utility in other service quality measures that were included to serve in building evidence of validity of the current instrument through their inclusion. There is research to suggest other dimensions of service quality in the healthcare setting that are not represented in the target behaviors presented by the institution. Five items developed uniquely for the medical setting by Winsted (2000) that represent the dimensions of Authenticity, Personalization, and Formality were included, as they seemed to have no overlap with the nine target service behaviors. Another two items were developed using Parasuraman, Zeithaml, and Berry’s (1988) dimension of Responsiveness, which is also not addressed by the institution but is documented to be critical to service quality in all sectors. The remaining two items were developed to measure perceptions of overall impact on patient experience and collaboration with colleagues. Using the nine service behaviors and drawing items from the service behavior literature to enhance validity, an initial pool of 46 items were generated for review.

To supplement the 46 rating scale items, eight behavior-specific multiple choice vignettes were developed to allow employees to provide the appropriate behaviors given caregiver-patient interaction situations. The vignettes were developed to evaluate employees’ practical applications of the Expected Service BehaviorsSM and assist in differentiation among employees based on their responses. The vignettes also serve to prompt employees to consider the expected behaviors by placing them in hypothetical,
conflicting situations in which they must select the most appropriate response. These items are presented in multiple-choice format, but differ from the rating scale items in that they could be scored as either correct or incorrect, based on the response chosen.

One final item was developed to serve to group respondents, given their unique roles and functions within the healthcare system. This question asked if the respondent, as part of their professional role, engages in direct patient care contact. The purpose of this item was to differentiate employees who more frequently engage with patients compared with those who interact more with visitors. It is relevant in analyzing if employees differ in their behaviors based on the nature of contact, whether it is primarily with patients or visitors. The addition of this item to the pool of 46 rating scale items and eight vignettes resulted in a pilot instrument consisting of 55 items.

The review and refinement process for the original items was assisted by third-party evaluation consultants to assess preliminary fit with the targeted service behaviors and clarity for administration. The evaluation experts are university-based program evaluators with over 30 years of experience in the health and human service arenas, which includes instrument development and data analysis. The evaluation consultants served as independent, external sources during the item generation and instrument development phase based on their extensive experience with similar projects. Upon consultation, the scope of the instrument was expanded to incorporate information about employee behaviors throughout their interactions with patients, colleagues, and patients’ visitors. These interactions were also assessed, as the institution’s expectation is that the service behaviors are expected to permeate all employee interactions, not just those
between patients and employees. A final review by the evaluation consultants and personnel from the target institution provided input on item wording, appropriateness, and overall impressions of the instrument prior to pilot testing administration.

For the purposes of this research, the construct of service quality was initially believed to be unidimensional, reflecting a single construct of high quality service. This assumption of unidimensionality, as well as any potential underlying factor structure, will be assessed using results from the pilot and subsequent administrations of the instrument and will be discussed in a later section. It is anticipated the results of the exploratory and confirmatory factor analyses will provide evidence of multiple dimensions, if any exist, prior to Rasch analysis. In other situations, assuming service quality as a single dimension may be problematic in defining the theoretical framework for item development, but in the current case service quality has been defined a priori and operationalized as the institution’s target service behaviors.

**The pilot instrument.** The instrument that was administered during the pilot testing phase consisted of 55 items. All items that were generated and deemed acceptable by the panel of experts were used in the pilot testing phase. Some redundant items were removed earlier in the item development phase based on feedback from the experts, while some were retained to serve as checks for validity and reliability. The instrument was designed so there were more items on the pilot than on the full-scale measure, allowing the data to indicate stronger items that were chosen to be retained on the final version.

The first item presented asked employees if they have direct patient care contact. This item was placed before the other items in case withdrawal or incompletion rates
were high. For each of the 46 rating scale items, employees were asked to provide ratings based on a five point Likert-type scale indicating the frequency of their behaviors (i.e., “1-Never or Almost Never” to “5-Always or Almost Always”). The Likert-type scale was applied two times within each item, where the employee was asked to provide distinct ratings that evaluate how (a) their patients or visitors, and (b) their colleagues would rate their frequency of target behaviors. Thus, employees are not asked explicitly if they engage in the behavior. Instead, the wording posits that the employee must provide a frequency rating for their behavior toward each of the two groups (i.e., patients/visitors and colleagues) in case they behave differently. For the eight vignettes, employees will be asked to select the most appropriate behavior given a set of four possible responses, with only one serving as the correct response. Since the instrument was administered within the employee’s first six months of service, the questions were worded to represent how they anticipate each group would provide the rating. A copy of the pilot instrument is included in Appendix A. The pilot instrument was expected to take approximately 15 minutes to complete and was done electronically using Qualtrics survey software.

Procedure

Data were collected by the institution and were provided for the purposes of this research de-identified, leaving only demographic information and instrument responses in the database to be used in the analysis. The pilot administration of the service behavior instrument took place within the first six months of service following the employees’ new-hire orientation. Employees were emailed electronic links to the online instrument.
on behalf of the institution’s Office of Patient Experience. The completed instruments were downloaded into a database that was used for analysis.

**Analyses**

The pilot data was analyzed using SPSS 21 (IBM, 2012). The vignette scores were recorded and reported separately from the rating-scale data, which is the focus of further analyses. First, pilot data was used to determine the reliability of the instrument, using an index of internal consistency (i.e., Coefficient Alpha). Secondly, an item analysis was conducted by analyzing item means, standard deviations, and inter-item correlations. This analysis was expected to yield information about internal validity as the newly-developed items were compared with those taken from the existing service behavior literature. Finally, an Exploratory Factor Analysis (EFA) was conducted to assist in identifying underlying factors present in the items. Identification of possible underlying factors provided some evidence of redundancy among items, which could support the use of a shorter measure. All analyses were approached through the lens of refining the instrument to make it as simple as possible without losing valuable information. A final version of the instrument was generated and supplied to the hospital for further use, with the analysis comprising the subsequent phase of the current research. The results of the pilot testing phase are presented in a later section.

**Refinement**

Based on the results of the pilot testing phase, some adjustments to the initial instrument were undertaken. An examination of item-level and test-level data was used to
evaluate the response options, reliability, and concurrent validity of the newly generated items with those from established instruments of service behavior. The results of the aforementioned evaluation are presented in a later section. Feedback from respondents was used to generate the final version, which was implemented during the full-scale administration.

Based on the results, but 16 of the rating scale items were identified as candidates for removal. Also, one vignette was met with over 90% correct response rate, so removal of the vignette was suggested. Based on the overall evaluation of the pilot results, some changes were made in order to make the instrument more user-friendly (i.e., shorter and involving less responses) and interpretable both in the research and in the practical contexts. A specific description of the refinement is presented in a following section.

**Part II: Full-Scale Administration**

The methodology of the full-scale administration mirrors that of the pilot testing phase, beginning with a discussion of participants and followed by a presentation of the measure, procedure, and analyses.

**Participants**

A total of 1,254 new participants were identified and sent electronic links to participate in the full-scale administration, requesting completion of the final version of the service behavior instrument. Participant selection and administration followed the same format as in the pilot testing phase, which was completed by the institution. Similar to the pilot testing phase, the revised service behavior instrument was administered
during employees’ first six months of service following their new-hire orientation.

However, the full-scale administration took place during a subsequent new-hire orientation to ensure participants in the pilot testing phase were not subject to a second administration; instead, they were new employees without previous exposure to the instrument. The full-scale administration took place over multiple orientations to accumulate enough participants to gather adequate sample data. The modality mirrored that of the pilot phase as well, where employees were e-mailed electronic links to the online instrument on behalf of the institution’s Office of Patient Experience. The completed instruments were downloaded into a database and sent to the researchers for use in the analysis.

Participants in the full-scale administration were also expected to represent a diverse range of demographic characteristics. The sample was anticipated to be fairly evenly distributed between genders and racial or ethnic backgrounds, as employees range from practitioners with advanced training and experience to general laborers who require minimal experience or positional qualifications. Differences in levels of education and career experience were present, where some positions included require only a high school education while others require undergraduate or graduate training. Some participants have prior experience in similar positions prior to their employment in the current setting.

Participant ages were predicted to cover a broad range, with most being from 21 to 65 years old. The inclusive nature of the data collection was predicted to provide a diverse sample of new employees, representative of various ages, racial or ethnic backgrounds, previous experience and levels of training.
The measure administered during the full-scale distribution was a refined version of the pilot instrument. Data from the pilot testing phase provided information that was useful in adjusting the measure as necessary to fulfill the needs of the institution while meeting criteria for reliability and validity. The intention of the pilot testing phase was to provide data relative to the quality of the items and of the scale as a whole, which was obtained and utilized. The pilot testing served to inform any needed changes and provide data for preliminary analyses prior to the full-scale administration, so the pilot instrument was refined prior to the full-scale administration.

The final measure consisted of 47 items. The retained items maintained the same structure, with one qualifying item evaluating the employee’s direct patient care contact, while 39 items employing the rating-scale format and the remaining seven of the vignette variety. The rating-scale items kept the same five-point Likert-type response options (i.e., “1-Never or Almost Never” to “5-Always or Almost Always”) utilized in the pilot instrument. The items were adjusted in their layout from the pilot instrument in which employees were asked to provide two prospective ratings for each of the behaviors based on how (a) their patients or visitors, and (b) their colleagues would likely indicate the frequency of their target behaviors. Instead, the full-scale instrument asked employees to provide only one self-rating, evaluating how patients or visitors would rate their behavior on 30 items and how their colleague would rate their behavior on the remaining nine items. The vignette items were also retained to elicit responses to hypothetical situations in multiple-choice format, with one response serving as the correct answer for each of the
seven items. The revised full-scale instrument was anticipated to take approximately 10 minutes to complete. A copy of the full-scale instrument is included in Appendix B.

**Procedure**

As in the pilot testing phase, data was collected by the institution and provided for the purposes of this research de-identified, leaving only demographic information and instrument responses in the database to be used in the analysis. Recruitment for participants to complete the full-scale service behavior instrument took place during a series of later employee new-hire orientations to ensure a different sample from the respondents included in the pilot testing phase. All eligible employees were given electronic links to the instrument to complete on behalf of the institution’s Office of Human Resources and all have been employed less than six months at the time of administration. The completed instruments were downloaded into a database that was used for analysis.

**Analyses**

The data collected during the full-scale administration was gathered within employees’ first six months of service and constituted the focal data employed in the analyses. The full-scale data were used in an attempt to further support the evidence of validity and reliability obtained in the pilot testing phase. An additional analysis of internal consistency (i.e., Coefficient Alpha) and individual item responses was done, and all reliability analyses took place using SPSS 21 (IBM, 2012). The vignette scores were recorded and reported separately from the rating-scale data, which is the focus of further analyses. The results of the full-scale administration are presented in a later section.
Based upon the factor structure established within the pilot testing phase, the results of the full-scale administration were also analyzed using Confirmatory Factor Analysis (CFA) techniques. The CFA was conducted using LISREL 9.10 (Joreskog & Sorbom, 2013). The results of the CFA were used to evaluate fit with the preliminary factor structure and to assess the underlying factor structure of the instrument. The intention of the factor analytic approaches in both the pilot testing phase and the full-scale administration is to generate and validate a factor structure using the nine service behaviors and the items developed from them. Factor structure, loading values, and inter-item correlations were examined to determine the dimensions of service quality captured by the items in the instrument and are presented in detail in the next chapter.

The final major analysis utilizing the full-scale rating data was an item analysis employing Rasch techniques. This analysis was completed using the Winsteps® 3.80.1 computer program (Linacre, 2013). The Rasch analysis, based upon Item Response Theory (IRT) principals, was expected to provide information on the range of difficulties presented by the items, the category usage based on the rating scale format, and any potential underlying factor structure. The Rasch model was used in place of other IRT models because it is the most parsimonious, appropriate model for the development of measures (Wright, 1997). The Rasch model fulfills the requirements of fundamental measurement, where more sophisticated IRT models fall short (Andrich, 2004). Incorporating Rasch analysis with the other instrument development analyses was expected to provide unique insights into the psychometric properties of the newly-developed instrument and into how the service behaviors are captured by the items.
Chief among the Rasch analysis includes data pertaining to item reliability, separation, fit, and category structure or thresholds. Each of these concepts provides insight into the instrument’s psychometric properties as part of the measure development process. In Rasch analysis, item reliability represents the ability to replicate item placement based on the Rasch estimates (Bond & Fox, 2007). Values closer to 1 represent strong reliability, while values closer to 0 indicate almost no certainty in replicating item difficulty estimates. Separation refers to the variation in item difficulties within the instrument, where larger values suggest good distribution of difficulties (de Ayala, 2009). Separation values that are less than one typically represent overlap or redundancy of items at a certain difficulty level. Reliability and separation are critical to consider, as are item fit statistics.

Rasch analysis also includes an examination of item and person fit statistics (i.e., infit and outfit) to address problematic responses. The infit and outfit analysis assist in identifying poor fit between the data and the Rasch measurement model. Infit and outfit statistics represent discrepancies between responses, with infit being weighted by values close to the expected value of difficulty or ability and outfit being unweighted, leading it to be more sensitive to outlying responses (de Ayala, 2009). It is essential that each item demonstrates adequate fit with the construct, as evidenced by fit statistics that represent correspondence with the Rasch model.

Additionally, the analysis considers the response option usage and category structure thresholds. The thresholds (i.e., step calibrations) are the difficulties estimated for choosing one response category over another (e.g., how difficult it is to endorse
“Always or Almost Always” instead of “Frequently”). The step calibrations should increase monotonically (i.e., have ascending threshold values), and the distance between threshold values should be neither too close together nor too far apart on the logit scale. Bond and Fox (2007) suggest thresholds should increase by at least 1.4 logits to show distinction between categories, but no more than 5 logits (i.e., to avoid large gaps in the variable). The thresholds will be analyzed to provide further evidence related to the category structure within the model and the use of response options on the instrument.

The assumption of unidimensionality of the instrument is a basic tenet of the Rasch model, and failure to evaluate this assumption may provide incorrect estimations and distorted results (Bond & Fox, 2007). Data from the Rasch analysis is anticipated to provide evidence of unidimensionality when the Rasch residuals are factor analyzed (Linacre, 1998). This analysis presented in the next chapter includes a factor analysis of the Rasch residuals to determine the dimensions present on the full-scale instrument. Further, the Rasch Factor Analysis (RFA) may not only determine if the assumption of unidimensionality was met, but it is also helpful in further identifying underlying constructs or components within the items (Linacre, 1998). The Rasch analysis was expanded to include a factor analysis of the Rasch model residuals to assist in determining dimensions within the service behaviors that are the focus of the instrument. Overall, analyses are expected to provide evidence of the psychometric properties and potential underlying dimensions of the new instrument.

Finally, the Rasch measurement approach was used to investigate Differential Item Functioning (DIF) and response invariance between employees with direct patient
care contact and employees without. The comparison of estimates across two or more
distinct groups of interest is the final step in investigating the quality of a new measure
(Bond & Fox, 2007). Employing DIF analysis is a direct test the invariance of items
between groups, in this case those with and without direct patient care contact.
Ultimately, DIF is a form of item bias in which significant differences are found between
groups of respondents (Lord, 1980). The examination of DIF determines if items have
different meanings for the different groups, which is evidence of bias within items. Items
that show evidence of DIF are problematic, and evidence of DIF in the development of a
new instrument would lead the items to be discarded or revised (Tennant & Pallant,
2007). Testing for item bias or DIF constitutes the last step in concluding the strength of
the newly-developed instrument.

The analyses conducted on the full-scale instrument (i.e., the CFA and Rasch
analysis) were anticipated to provide different, but useful information in drawing
conclusions related to the individual items and the instrument as a whole. The purpose of
the confirmatory analysis was to determine how items align with one another. Namely, it
was expected that the items will be responded to similarly, with each item dependent on
the behavior from which it was developed. Ultimately, the CFA was included to provide
evidence of how well the expected behaviors are portrayed by their component items.
The Rasch analysis, on the other hand, was utilized to determine if the items collectively
belong to one dimension, that representing overall service quality. The different analyses
were expected to provide separate, unique results that will help draw conclusions about
whether the behaviors can function independently and how well they can explain good, quality service.

Overall, analyses were expected to yield evidence of the psychometric properties and potential underlying dimensions of the new instrument. The processes of evaluating these characteristics using basic item analysis and EFA for the pilot phase were used to generate the refined full-scale instrument. Evaluating the basic item analysis, CFA, and Rasch analysis of the full-scale instrument was expected to provide evidence of the strength and utility of the newly-developed assessment. The former explanations of the methods were provided to serve as a framework for a critical evaluation of the results. Both the results of the pilot and full-scale administration phases are presented in detail in the following chapter.
CHAPTER IV

ANALYSIS OF THE RESULTS

As in the explanation of the methodology, the results are presented in two major sections; the first outlines the analyses and results of the pilot testing while the latter outlines those of the full-scale administration.

Part I: Pilot Testing Phase

The pilot administration analyses included an item analysis followed by an EFA as a process for data reduction. Between the pilot and full-scale administrations, steps taken to refine the instrument are described in detail.

Sample

The initial request for participation in the pilot phase was sent via company e-mail to 570 newly-hired employees. Reminders to non-responders were sent weekly from the institution’s Office of Patient Experience for the nine weeks the instrument was open for data collection. Of the original sample of 570 employees, 165 provided consent and began the pilot instrument (28.95%). However, the number of completed instruments was only 68, comprising 11.92% of the entire sample (Table 2). While this was a targeted sample with participation requests coming from an official office within the institution, the response rate lags behind the expectation using the online survey method (Nulty, 2008). Table 2 presents the pilot instrument completion percentage for the sample.
Table 2.

<table>
<thead>
<tr>
<th>Percent of Items Complete</th>
<th># of Instruments</th>
<th>% of Instruments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-10%</td>
<td>45</td>
<td>27.27</td>
</tr>
<tr>
<td>11-20%</td>
<td>15</td>
<td>9.09</td>
</tr>
<tr>
<td>21-30%</td>
<td>2</td>
<td>1.21</td>
</tr>
<tr>
<td>31-40%</td>
<td>9</td>
<td>5.45</td>
</tr>
<tr>
<td>41-50%</td>
<td>3</td>
<td>1.81</td>
</tr>
<tr>
<td>51-60%</td>
<td>10</td>
<td>6.06</td>
</tr>
<tr>
<td>61-70%</td>
<td>2</td>
<td>1.21</td>
</tr>
<tr>
<td>71-80%</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>81-90%</td>
<td>6</td>
<td>3.63</td>
</tr>
<tr>
<td>91-100%</td>
<td>73</td>
<td>44.24</td>
</tr>
</tbody>
</table>

Note. Percentage of instruments column is based on the 165 respondents that provided consent.

The instrument remained open for data collection over a period of nine weeks, and considering weekly reminders were sent to non-responders, the completed response rate hovering just under 12% is worrisome. The data in Table 2 is provided to assist in illustrating that a substantial portion of respondents (37.57%) did not make it through one-third of the instrument. Based on that information, a suggested revision for the full-scale administration was to adapt the items to reduce the amount of time necessary to complete the entire instrument. The subsequent analyses are based on the full sample size of 165 respondents given the responses they provided; thus, frequencies vary based on the items completed by each individual.

A majority of the employees that participated in pilot data collection were female (n = 140, 84.84%), while gender had no association with 90% or more of the items complete ($\chi^2 (1) = 1.79, p > .05$). The cross-tabulation between gender and completion is
provided in Table 3. While males were under-represented within the sample, their completion rates were not substantially different from females.

Table 3.

<table>
<thead>
<tr>
<th>Gender</th>
<th>Complete Instruments (%)</th>
<th>Incomplete Instruments (%)</th>
<th>Total (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>65 (39.39)</td>
<td>75 (45.45)</td>
<td>140 (84.84)</td>
</tr>
<tr>
<td>Male</td>
<td>8 (4.84)</td>
<td>17 (10.30)</td>
<td>25 (15.15)</td>
</tr>
<tr>
<td>Total</td>
<td>73 (44.24)</td>
<td>92 (55.75)</td>
<td>165 (100)</td>
</tr>
</tbody>
</table>

*Note.* Complete instruments signify responses collected for more than 90% of items.

When respondents were asked if they engaged in direct patient care contact, responses were provided by 151 employees. Most of the employees (66.88%) indicate their current role involves direct patient care contact.

**Item Analysis**

The first step in item analysis was to reverse code items that are negatively-worded. The pilot instrument contained two items that could be interpreted as negatively-worded: “I take patients’/visitors’ complaints personally” and “If I’m having a challenging day, patients/visitors can tell.” The responses for these items were reversed so lower scores indicate poor service quality or negative service behaviors. With the reversed scoring for those items, now the instrument can be analyzed to represent high item and total scale scores reflecting high quality service and exhibition of expected behaviors.

Coefficient Alpha was first calculated using all items together instead of separating potential scales due to the assumed underlying factor of “service quality”. Based upon the 68 completed instruments, the 46 rating scale items demonstrate strong
internal consistency ($\alpha = .907$) when attributed to patients or visitors, as well as when attributed to colleagues ($\alpha = .905$). These estimates of internal consistency represent the average of all possible split-half coefficients when dividing the test in half (Cronbach, 1951), and they are universally considered to be high. Such an elevated Coefficient Alpha could be indicative of the items representing a single, unidimensional factor (e.g., service quality), or could be a function of a large pool of items or homogeneity of the tested group (Crocker & Algina, 2008). To determine the nature of the inter-relationships among the items, the investigation beyond internal consistency ventures toward item statistics.

The items were rated by participants on a five-point Likert-type scale, with values ranging from one to five. When referring to patients or visitors, item means ranged from 2.62 (i.e., Item 3: “I take patients’/visitors’ complaints personally.”) to 4.97 (i.e., Item 9: “I treat patients with care.”). Most item means fell above 4.00, suggesting some negative skew as items were rated highly by the employees. When referring to colleagues, item means ranged from 2.69 (i.e., Item 3: “I take patients’/visitors’ complaints personally.”) to 4.98 (i.e., Item 9: “I treat patients with care.”). Ratings were very similar in reference to both patients/visitors and colleagues, where colleague means were also predominantly values of 4.00 or higher. Overall, employees report they engage in these behaviors at least frequently when interacting with patients, visitors, and their colleagues.

Potential threats to internal consistency are identified when examining the values for the Coefficient Alpha statistic if item deleted (Table 4). This represents what value Coefficient Alpha would assume if the item was removed from the analysis (Dimitrov,
2009). Items that are considered problematic to reliability are those with coefficient alpha values higher with the item excluded than the obtained alpha with the item included. The results indicate that Item 3 (i.e., “I take patients’/visitors’ complaints personally.”) poses a threat to internal consistency, with Alpha if Item Removed of .923 for both patients/visitors and colleagues. Item 7 (i.e., “If I'm having a challenging day, patients/visitors can tell.”) was also found to be problematic, as the Alpha if ItemRemoved value matches the patient/visitor total alpha ($\alpha = .907$) but exceeds the colleague total scale alpha of .905. Item mean, standard deviation, and Alpha if ItemRemoved statistics are presented in Table 4.
### Table 4.

**Descriptive Statistics and Alpha if Item Deleted for Items on the Pilot Instrument**

<table>
<thead>
<tr>
<th>Item</th>
<th>Ratings for Patients/Visitors</th>
<th>Ratings for Colleagues</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>1.)</td>
<td>I am knowledgeable about patient care.</td>
<td>4.540</td>
</tr>
<tr>
<td>2.)</td>
<td>I acknowledge visitors and welcome them.</td>
<td>4.750</td>
</tr>
<tr>
<td>3.)</td>
<td>I take patients'/visitors' complaints personally.</td>
<td>2.660</td>
</tr>
<tr>
<td>4.)</td>
<td>I introduce myself when I meet patients/visitors.</td>
<td>4.770</td>
</tr>
<tr>
<td>5.)</td>
<td>I make patients/visitors feel comfortable.</td>
<td>4.820</td>
</tr>
<tr>
<td>6.)</td>
<td>I use patients'/visitors' preferred name when I speak with them.</td>
<td>4.630</td>
</tr>
<tr>
<td>7.)</td>
<td>If I'm having a challenging day, patients/visitors can tell.</td>
<td>4.030</td>
</tr>
<tr>
<td>8.)</td>
<td>I am genuine when interacting with patients/visitors.</td>
<td>4.900</td>
</tr>
<tr>
<td>9.)</td>
<td>I treat patients with care.</td>
<td>4.960</td>
</tr>
<tr>
<td>10.)</td>
<td>Each interaction I have with patients/visitors is authentic and individualized.</td>
<td>4.750</td>
</tr>
<tr>
<td>11.)</td>
<td>I summarize back to patients/visitors what they communicate with me.</td>
<td>4.310</td>
</tr>
<tr>
<td>12.)</td>
<td>I can follow through on my responsibilities and still meet patients' expectations.</td>
<td>4.650</td>
</tr>
<tr>
<td>13.)</td>
<td>I encourage patients/visitors to discuss or present any concerns throughout their visit.</td>
<td>4.490</td>
</tr>
<tr>
<td>14.)</td>
<td>&quot;Polite&quot; and &quot;thoughtful&quot; are two words patients/visitors use to describe me.</td>
<td>4.780</td>
</tr>
</tbody>
</table>
Table 4. (continued)

<table>
<thead>
<tr>
<th>Item</th>
<th>Ratings for Patients/Visitors</th>
<th>Ratings for Colleagues</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>15.)</td>
<td>4.390</td>
<td>0.580</td>
</tr>
<tr>
<td>16.)</td>
<td>4.710</td>
<td>0.532</td>
</tr>
<tr>
<td>17.)</td>
<td>4.730</td>
<td>0.512</td>
</tr>
<tr>
<td>18.)</td>
<td>4.770</td>
<td>0.426</td>
</tr>
<tr>
<td>19.)</td>
<td>4.390</td>
<td>0.707</td>
</tr>
<tr>
<td>20.)</td>
<td>4.910</td>
<td>0.281</td>
</tr>
<tr>
<td>21.)</td>
<td>4.520</td>
<td>0.813</td>
</tr>
<tr>
<td>22.)</td>
<td>4.760</td>
<td>0.432</td>
</tr>
<tr>
<td>23.)</td>
<td>4.950</td>
<td>0.226</td>
</tr>
<tr>
<td>24.)</td>
<td>4.860</td>
<td>0.404</td>
</tr>
<tr>
<td>25.)</td>
<td>4.290</td>
<td>1.110</td>
</tr>
<tr>
<td>26.)</td>
<td>4.580</td>
<td>0.587</td>
</tr>
<tr>
<td>27.)</td>
<td>4.820</td>
<td>0.415</td>
</tr>
<tr>
<td>28.)</td>
<td>4.540</td>
<td>0.786</td>
</tr>
<tr>
<td>Item</td>
<td>Ratings for Patients/Visitors</td>
<td>Ratings for Colleagues</td>
</tr>
<tr>
<td>------</td>
<td>------------------------------</td>
<td>------------------------</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>29.)</td>
<td>I ask others how they prefer I address them (i.e. what they like to be called).</td>
<td>4.070</td>
</tr>
<tr>
<td>30.)</td>
<td>I am timely in my response to patients'/visitors' concerns or requests.</td>
<td>4.670</td>
</tr>
<tr>
<td>31.)</td>
<td>I make eye contact when speaking with patients/visitors.</td>
<td>4.910</td>
</tr>
<tr>
<td>32.)</td>
<td>I explain my title or role.</td>
<td>4.600</td>
</tr>
<tr>
<td>33.)</td>
<td>I greet others warmly.</td>
<td>4.900</td>
</tr>
<tr>
<td>34.)</td>
<td>I do my best to assure patients'/visitors' needs are met.</td>
<td>4.880</td>
</tr>
<tr>
<td>35.)</td>
<td>I am mindful of patients' privacy and dignity.</td>
<td>4.860</td>
</tr>
<tr>
<td>36.)</td>
<td>I respond with appropriate emotions when interacting with patients/visitors.</td>
<td>4.820</td>
</tr>
<tr>
<td>37.)</td>
<td>I am sincere in my apologies.</td>
<td>4.900</td>
</tr>
<tr>
<td>38.)</td>
<td>I acknowledge and respect patients'/visitors' emotions.</td>
<td>4.890</td>
</tr>
<tr>
<td>39.)</td>
<td>I do my best to show compassion for patients or families.</td>
<td>4.900</td>
</tr>
<tr>
<td>40.)</td>
<td>I encourage patients/visitors to ask for help when necessary.</td>
<td>4.820</td>
</tr>
<tr>
<td>41.)</td>
<td>When I forward a patient's/visitor's concern, I follow up to ensure it has been handled.</td>
<td>4.370</td>
</tr>
<tr>
<td>42.)</td>
<td>When I apologize, I take care not to place blame.</td>
<td>4.700</td>
</tr>
</tbody>
</table>
Table 4. (continued)

<table>
<thead>
<tr>
<th>Item</th>
<th>Ratings for Patients/Visitors</th>
<th>Ratings for Colleagues</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>43.)</td>
<td>I can clarify or explain what patients/visitors do not understand.</td>
<td>4.420</td>
</tr>
<tr>
<td>44.)</td>
<td>I strive to be professional in my interactions.</td>
<td>4.940</td>
</tr>
<tr>
<td>45.)</td>
<td>I recognize that each patient/visitor has a unique experience and circumstances.</td>
<td>4.860</td>
</tr>
<tr>
<td>46.)</td>
<td>I close every interaction by saying &quot;Thank you&quot; to patients/visitors.</td>
<td>4.660</td>
</tr>
</tbody>
</table>

*Note.* Coefficient alpha was calculated separately for patients/visitors (α = .907) and for colleague ratings (α = .905).
**Correlations.** Inter-item correlations were calculated to determine the nature of the relationship between items. The Pearson $r$ statistic is not applicable due to the ordinal nature of the data collected, thus, Spearman’s $\rho$ was the most appropriate method in determining the correlation among item scores (Hienkle, Wiersma, & Jurs, 2003). The $\rho$ values were calculated for the items in three separate analyses, once each for the patient/visitor ratings, the colleague ratings, and for the relationship between patient/visitor and colleague ratings. For the separate correlations of patient/visitor and colleague ratings, all items were run to analyze their inter-relationship, or how each of the 46 items relate with all of the others. To examine the relationship between patient/visitor and colleague ratings, the correlations were calculated to define the relationship between the ratings on matched items (i.e., how employees rated themselves in terms of patients/visitors and colleagues on the same item). The correlations between patient/visitor and colleague items are included in Appendix C.

For the patient/visitor ratings, the lowest Spearman’s $\rho$ value (-.277) was obtained between Item 30 (i.e., “I am timely in my response to patients'/visitors' concerns or requests.”) and Item 3 (i.e., “I take patients'/visitors' complaints personally.”), even after the latter was reverse-coded. This is the strongest inverse relationship, meaning that employees were likely to rate one more highly and the other low. Negative relationships were found in 80 pairs of items, with many of them weak relationships obtaining values close to 0. The strongest positive relationship was found between Item 30 and Item 26 (i.e., “I am prompt in my service delivery.”), with a Spearman rho of .704. This reflects employees very likely to rate both items very similar with each other.
Together, most inter-item correlations were rather weak (i.e., less than .400). However, 44 of the 46 items demonstrated at least a moderate relationship (i.e., .400 or higher) with at least one other item. The problematic item was Item 3 (i.e., “I take patients'/visitors' complaints personally.”), which showed a very weak and negative relationship with most items. This could likely be due to some inherent confusion in item wording or responses associated with this behavior. Similarly, Item 7 (i.e., “If I'm having a challenging day, patients/visitors can tell.”) was somewhat troublesome as it did not demonstrate at least a moderately strong relationship with any other items, but correlation values were frequently .300 and higher. This item was also a candidate for removal when the instrument was revised.

When employees rated their behaviors from a colleague’s point of view, the results shared much in common with the patient/visitor ratings. Most of the correlations were weak and positive, with some evidence of inverse relationships among items. The strongest negative correlation (-.282) was obtained between Item 15 (i.e., “I anticipate the needs of my patients/visitors.”) and Item 3. Item 3 continued to have a negative relationship with most other items, while Item 9 (i.e., “I treat patients with care.”) also had a substantial amount of negative correlations with other items. Items 3 and 9 were the only two which failed to achieve at least a moderate correlation with any other item. The strongest positive relationship (.690) was found between Item 39 (i.e., “I do my best to show compassion for patients or families.”) and Item 37 (i.e., “I am sincere in my apologies.”). Such strong positive correlations may provide evidence of redundancy within the items, which will be examined in the refinement phase.
The correlations between patient/visitor and colleague ratings on the same item were also calculated and analyzed. Here, high correlations indicate the employee perceives his or her behavior to be similar whether interacting with patients/visitors or colleagues. These groupings were purposefully split in instrument development in hopes of collecting data sensitive to the differences in interactions between coworkers and customers. The results indicate the employees self-reported very similar behavioral frequencies whether interacting with patients/visitors or colleagues. The obtained rho values ranged from .336 (Item 9: “I treat patients with care.”) up to 1.000 (Item 31: “I make eye contact when speaking with patients/visitors.”), meaning for that item the patient/visitor and colleague ratings were identical for all respondents. In the case of Item 9, employees provided ratings that were substantially different whether referring to patients/visitors or their colleagues. Most correlations between patient/visitor and colleague ratings were very strong (i.e., .800 and above), suggesting employees consistently display these behaviors regardless of the audience. This may also provide evidence of redundancy in questioning, where ratings are so similar that a single rating could be used to account for both interactions.

The last investigation employing item correlations was to examine content and construct validity of the new instrument through the items that have been used in previous service quality literature. This was achieved through evaluation of items that have demonstrated utility in other measures of service quality (Parasuraman, Zeithaml, & Berry, 1988; Winsted, 2000) compared with the newly-developed items specific to the institution’s target behaviors. Specifically, the five items adapted from Winsted (2000)
and the two items from Parasuraman, Zeithaml, and Berry (1988) were examined for their correlations with the other items using item-total correlation and inter-item correlation values (Table 5). It is expected that the seven target items would demonstrate positive correlations with the newly-developed items and item-total correlations. These items serve two purposes: a) to supplement the items based on the institution’s expected behaviors that do not encompass behaviors that reflect high quality service, and b) to provide a means of comparison between quality service behaviors and the items that have been developed. The dual purpose behind the inclusion of these items leads away from expecting very strong correlations. The item-total correlations for both patient/visitor and colleague ratings are presented in Table 5.

Table 5.

<table>
<thead>
<tr>
<th>Item</th>
<th>Patient/Visitor</th>
<th>Colleague</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.)</td>
<td>I am knowledgeable about patient care.</td>
<td>0.230</td>
</tr>
<tr>
<td>10.)</td>
<td>Each interaction I have with patients/visitors is authentic and individualized.</td>
<td>0.559</td>
</tr>
<tr>
<td>15.)</td>
<td>I anticipate the needs of my patients/visitors.</td>
<td>0.635</td>
</tr>
<tr>
<td>26.)</td>
<td>I am prompt in my service delivery.</td>
<td>0.357</td>
</tr>
<tr>
<td>30.)</td>
<td>I am timely in my response to patients'/visitors' concerns or requests.</td>
<td>0.332</td>
</tr>
<tr>
<td>44.)</td>
<td>I strive to be professional in my interactions.</td>
<td>0.247</td>
</tr>
<tr>
<td>45.)</td>
<td>I recognize that each patient/visitor has a unique experience and circumstances.</td>
<td>0.610</td>
</tr>
</tbody>
</table>

*Note.* Items 26 and 30 were adapted from Parasuraman, Zeithaml, & Berry (1988). The remaining items were adapted from Winsted (2000).

The calculated item-total correlations suggest that these items demonstrate fit with the newly-developed items, as the values ranged from .230 to .635 when prompted to provide patient/visitor ratings, and from .367 to .678 when prompted to provide colleague.
ratings. The elevated correlations found for items 10, 15, and 45 represent a very strong association between these items and the scale total score. The lower values obtained are not necessarily problematic, as they were included to supplement previously-defined quality service behaviors not encompassed by the institution’s expectations. Further, an examination of the inter-item correlations between these items and the others provided no evidence of problematic items or relationships. There were no unusual patterns our outlying correlation values that make these adaptations distinct from the remaining items, suggesting fit with the construct of service quality and agreement between the items. Overall, the results provide evidence of construct validity while highlighting potential redundancy among the items.

**Exploratory Factor Analysis of the 46 Rating-Scale Items**

An exploratory factor analysis using principal axis factoring (PAF) and oblique (Direct Oblimin) rotation was conducted on the patient/visitor rating data to determine potential underlying dimensions within the items. In the first factor analysis, all items were included with the exception of Item 3, given the concerns presented in the item analysis. The ordinal nature of the rating-scale data called for the factor analysis to be conducted using Spearman’s rho coefficients instead of Pearson’s $r$. The use of Spearman rho coefficients is more appropriate given the extraction of factors from a matrix of associations among the items (Gorsuch, 1983). Due to the high rate incomplete instruments, the usable sample for the EFA was 68. This is considered less than optimal sample size given the number of items, as results are questionable when conducted with samples less than 100 (Gorsuch, 1983) or when a ratio of participants to items of 20:1 is
not achieved (Hair, Anderson, Tatham, & Black, 1995). Patient/visitor ratings were used as they were the more complete dataset, allowing for a larger sample size than could be achieved using the colleague ratings.

The Kaiser-Meyer-Olkin (KMO) measure of sample adequacy and Bartlett’s test of sphericity were used to evaluate the appropriateness of the factor analysis based on the inter-item correlation matrix. KMO is a standardized index representing the strength of each item’s relationship with the other items in the matrix, ranging from 0 to 1 (Pett et al., 2003). The items must demonstrate at least a moderate relationship with each other for the factor analysis to be considered (Kaiser, 1974). The data suggest an adequate relationship among items was achieved with a moderate KMO value of .639. Bartlett’s test of sphericity was calculated to evaluate if the correlation matrix obtained is an identity matrix, which would indicate no relationship among items (Bartlett, 1950). Bartlett’s test was statistically significant ($\chi^2_B = 2226.450, df = 990, p < .001$) indicating the correlation matrix is not an identity matrix and factoring is appropriate to pursue given the data.

The scree plot, which represents the eigenvalues prior to extraction and rotation, visually supports the extraction of about seven factors (Cattell, 1966). To verify the extraction of factors beyond examination of the scree plot, Kaiser’s criterion was adopted to determine factors with eigenvalues greater than one (Guttman, 1954). The Direct Oblimin procedure (Jennrich & Sampson, 1966) was applied to rotate the factors (i.e., $\delta = 0$) to better approximate simple structure and a more meaningful factor solution (Thurstone, 1947). This oblique rotation was selected over orthogonal methods due to the
anticipated correlation among the factors (Thompson, 2004). Twelve factors obtained eigenvalues greater than one, with the sums of squared loadings from these components having a cumulative value of 66.30% in explaining the total variance within the data.

The structure matrix represents the correlations of the items with each of the rotated factors. Each of the items demonstrated a moderate relationship with at least one of the factors with a correlation of .400 or higher, except Item 32 (i.e., “I explain my title or role.”). In many cases, an item would have a structure coefficient value exceeding .400 with multiple factors, providing further evidence of relationships among the factors. Analysis of the pattern matrix, which presents the partial regression coefficients for items within factors, indicated most items were likely to have a unique relationship with a factor while controlling for the correlation among factors. At least two items were identified for each of the 12 factors with pattern coefficients greater than .300, while many factors contained multiple items exceeding .400. The pattern and structure matrices combine to provide insight into item alignment within the newly-identified factors.

An examination of the factor correlation matrix indicated that the 12 factors demonstrated little statistical relationship with one another. Factor correlations ranged from .313 to -.294, suggesting there was minimal overlap among items and the relationship between factors was weak. Finally, Coefficient Alpha was calculated for each of the factors’ component items. Alpha values ranged from .251 (Factor 3) to .794 (Factor 6). The internal consistency within factors tended to be low when one or more items had a negative pattern coefficient (i.e., Factors 2, 3, and 11) or there were only two
items defining the factor (i.e., Factor 12). Together, the data indicate the items can be represented by their underlying factors which are distinct from one another.

The identification of 12 factors from 46 items led to difficulty in defining the factors based on constituent items. While a factor structure was established, some items did not accrue the strength in defining the factors that was expected. As more factors were identified, multiple items became less apt to fit together into new dimensions. Several of the pattern coefficients fell between .300 and .400, while others were quite strong (i.e., .700 and above). A more parsimonious structure was sought both to reduce the number of items on the full-scale instrument and to enhance the likelihood of confirming the factor structure in future analyses. The results of the EFA were used to select items as candidates for removal; namely, those with insufficient loadings on the pattern matrix or those with multiple loadings on the structure matrix.

All items achieved a loading greater than .300 on at least one factor, so no items could be identified based on a common definition of a weak loading (Hair, et al., 1995). In order to further simplify factor structure, items with weaker factor loadings were examined and considered for removal. Based on the lower loadings, these items were not strong indicators of the factors to which they ascribed statistically. Given the different number of items aligning with each factor, items with weaker loadings were considered for removal based on the number of other items within that factor. Using pattern matrix values, six items were identified as having weak factor loadings (i.e., Items 10, 19, 22, 23, 38, and 43) and were removed from the analysis. Other items (i.e., Items 21, 28, and
31) were stronger representations of their respective factors and provided a more parsimonious explanation of the factors.

Items that were determined problematic or worthy of investigation for removal were those that fit one of three conditions: a) they demonstrated weak loadings on the pattern matrix suggesting they did not strongly align with one factor, b) items which load on more than one factor, or c) items with negative or inverse pattern coefficients. None of the items exhibited loadings exceeding .400 on more than one factor after rotation, so no items were identified as problematic using that criterion. Several items did provide loadings that were negative: Item 32 (-.351, Factor 2), Item 11 (-.459, Factor 3), and Item 4 (-.338, Factor 11) obtained factor loadings opposite the others within their respective factors. Thus, the six items with lower factor loadings (i.e., Items 10, 19, 22, 23, 38, and 43) the three items (i.e., Items 32, 11, and 4) with inverse loadings were further investigated in the refinement phase.

**Refinement**

The goal of the refinement phase was to reduce the number of items to a more manageable level by eliminating those that were problematic or redundant using available data provided by the results from the item analysis and EFA. Given the poor completion percentage of the pilot instrument (11.92%), a streamlined and more efficient full-scale instrument was expected to elicit a much stronger response. While reducing the number of items and subsequent length of time to completion, it was imperative not to substantially impact the quality of the measure outlined through the analysis of the pilot instrument. The following provides a description of the factors that went into applying
the results of the pilot phase to development of a much less time-intensive version, applied during the full-scale administration.

The data suggest the total number of responses needed to be reduced, where 44.83% of respondents completed less than one half of the instrument. Though items were separated into paginated groups so as to not burden participants, the sheer amount of ratings employees were asked to provide was likely interpreted as overwhelming. The first step in refinement was to drastically shorten the amount of responses the participants were requested to provide, with the intention of maintaining quality over quantity. The pilot instrument consists of 46 rating-scale items, each endorsed twice by respondents to reflect how a) patients/visitors and b) colleagues would rate their behavior. Thus, the total number of rating-scale responses collected in the pilot phase was 92. Combining the rating-scale responses with the 8 vignettes, participants were asked to provide answers to 100 questions, not including whether or not they had direct patient care contact as part of their position.

The first logical step in reducing the number of responses was to examine the unique ratings employees were asked to provide. The relationship between patient/visitor ratings and colleague ratings provided by the same employee was very strong (i.e., Spearman’s rho greater than .900 for most items), suggesting redundancy in asking employees to provide the distinct ratings. Only one item, Item 9, was found to demonstrate a moderate relationship (i.e., Spearman’s rho = .336) between patient/visitor and colleague ratings, while all other rho values exceeded .780. Reducing the number of items in half by asking respondents to provide only one rating (i.e., “Patients/Visitors
would say that…”) is considered more feasible since pilot data indicate employee behaviors are consistent regardless of with whom they are interacting. The adjustment of reducing dual-ratings for all items would allow for data collection to take much less time than when employees were asked to provide two distinct ratings for each item.

The unique case of Item 9, where the patient/visitor and colleague ratings were not similar, led to the item being asked twice on the full-scale instrument. The stem “Colleagues would say that…” was used in conjunction with a small group of other items; namely, those that were questionable when evaluated for removal. Nine items were selected to be presented in this section, only four of which were redundant with the patient/visitor rating section. These items were those for which more data was needed to determine their fit and utility among the other items on the instrument. This alternate section was presented following the vignettes, so if participants decided to withdraw in the middle of the instrument, the most useful items were presented.

The decision sequence for item removal began with the item analysis results, followed by consideration for the fit of items into components through EFA, and concluding with a critical review of item content for redundancy. The data from the individual items were examined through the lens of keeping only the most useful, appropriate, and sound items for use in the full-scale instrument. Based on increases in Coefficient Alpha if the items were removed, Item 3 and Item 7 were strong candidates for removal. This was likely due to the negative wording included in the items, making it confusing for employees to provide accurate responses. Subsequent investigations were necessary to generate more items with a sound basis for removal.
In principal, the pilot instrument was anticipated to include redundancy within items. Given the nature of the initial development using behavioral indicators, items constituting the same indicator were expected to demonstrate similarity with one another. Determining repetition of items was concluded through two methods: a) review of the EFA results for dimensions with a large number of items, and b) grouping items based on the service behavior which they reflect. Together, the items that were included on the full-scale instrument were expected to uphold a similar factor structure while representing the foundational expected behaviors of quality service. The factor structure obtained in the pilot phase was subjected to confirmation using the full-scale administration data; however, a representation of each of the Expected Service Behaviors was necessary to be retained.

Applying the results of the EFA, items posed as contenders for removal were those with low factor loadings on the pattern matrix (i.e., pattern coefficients) and those with negative or inverse loadings when compared to other items within their factors. Pattern coefficients that are not sufficiently large (i.e., .300 or higher) suggest the effect of the factor on the item, when other factors are controlled for, is not strong. Coefficients between .300 and .400 on the pattern matrix were found for eight items (i.e., Items 4, 7, 18, 19, 23, 27, 32, and 38). These pattern coefficients served as a starting place for item decision-making and did not define which items were retained; the values presented in the structure matrix were considered as well. Also, items which posited negative loadings were identified (i.e., Items 4, 11, and 32) and evaluated for their contribution to the instrument. Only Item 11 (i.e., “I summarize back to patients/visitors what they
communicate with me.”) had a moderately strong negative loading (-.459) and was removed. Overall, priority in selection of items for removal went to review of the content of the item after all data were analyzed.

Weighing all available data from the EFA and following critical examination of item content for behavioral indicators, two criteria were used to determine recommendations for removal (Thompson, 2004). First, dimensions that were newly-identified within the EFA were used to pare items from factors that consisted of multiple strongly-related items. Concurrently, redundancy and candidacy for removal were identified when examining the number of items constituting each of the Expected Service BehaviorsSM. Thus, items that most suggested unnecessary repetition were those that occurred within larger dimensions based on the EFA, and those referencing a behavioral indicator represented in many other items. Items that served as the lone representation of a behavioral indicator or of a new dimension were most likely to be retained.

To evaluate the performance of items based on the behavioral indicator they represent, Coefficient Alpha was calculated for each of the groupings for which there was more than one item (Table 6). As anticipated, lower values for Coefficient Alpha were found for scales with fewer items, and the largest value (i.e., $\alpha = .759$) was found for the Empathy scale, which consisted of seven items. The eighth behavioral indicator, using common courtesy, was presented in eight items, where two of which have already been identified as problematic (i.e., Items 3 and 7). Excluding the two problematic items, the courtesy scale had moderate internal consistency ($\alpha = .645$). The numbers of items
representing each indicator and calculated values for Coefficient Alpha are provided in Table 6.

### Table 6.

**Internal Consistency within Item Groupings by Behavioral Indicators**

<table>
<thead>
<tr>
<th>Behavioral Indicator</th>
<th>Items</th>
<th>Coefficient Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.) Acknowledge the other person</td>
<td>1</td>
<td>---</td>
</tr>
<tr>
<td>2.) Introduce self and role</td>
<td>2</td>
<td>0.546</td>
</tr>
<tr>
<td>3.) Use person's preferred name and greet warmly</td>
<td>3</td>
<td>0.470</td>
</tr>
<tr>
<td>4.) Clearly communicate expectations</td>
<td>2</td>
<td>0.358</td>
</tr>
<tr>
<td>5.) Offer to resolve concerns or forward to the right person</td>
<td>5</td>
<td>0.612</td>
</tr>
<tr>
<td>6.) Use active listening</td>
<td>5</td>
<td>0.494</td>
</tr>
<tr>
<td>7.) Show empathy</td>
<td>7</td>
<td>0.759</td>
</tr>
<tr>
<td>8.) Use common courtesy*</td>
<td>7</td>
<td>0.645</td>
</tr>
<tr>
<td>9.) Offer to help</td>
<td>3</td>
<td>0.520</td>
</tr>
</tbody>
</table>

*Note. Coefficient Alpha was not computed for the lone item of behavioral indicator 1. *Coefficient Alpha was computed with problematic items removed

Given the limited sample size and some behavioral indicators being represented by only two items, the calculated values for Coefficient Alpha are not sufficient evidence of redundancy to be able to remove items from the scale. The obtained values for Coefficient Alpha were found to be moderate, at best, for a few of the behavioral indicators (i.e., Indicators 5, 7, and 8). The remaining item groupings obtained Coefficient Alpha values that were low or considered unacceptable. Such low values for Coefficient Alpha suggest a weak relationship among items and poor internal consistency within the scales.

An investigation into inter-item correlations within these groupings was also undertaken. Examination of the correlations between items that represent the same
behavioral indicator provided insight into which items best fit together to portray the behavior of interest. Correlations between items representing the same behavior were expected to be high; suggesting redundancy in the behavioral indicator that was presented. However, just as problematic are items with very low correlations (i.e., .200 or below) with other items in the same group or those with negative correlations. Thus, the correlations were used to identify both items of over-represented behaviors through high correlations, and items that do not adequately represent the behavioral indicator through low or negative correlations. Candidates for removal were identified through a critical examination of these item inter-relationships.

A total of 16 items were identified as candidates for removal following critical examination. Item 3 was removed after identification as problematic based on a higher Alpha if item removed, while nine items (i.e., Items 5, 11, 19, 22, 29, 30, 36, 38, and 43) were removed which demonstrated redundancy but poor fit within behavioral indicators. Items 19, 22, 38, and 43 were also met with weak loadings within their dimensions identified by the EFA, while Item 11 was removed due to a negative pattern coefficient. Item 42 (i.e., “When I apologize, I take care not to place blame.”) was also removed due to potential confusion in item wording, combined with difficulty in determining corresponding behavioral indicator and misalignment with the other items in its newly-identified dimension. Of the 16 items identified as candidates for removal, 11 were removed entirely (i.e., Items 3, 5, 11, 19, 22, 29, 30, 36, 38, 42, and 43). The remaining five items, Items 7, 10, 15, 23, and 35, were removed from the patient/visitor section but retained in the section where colleague ratings were requested because of reasonable
factor loadings within their respective dimensions based on the EFA, though they were reflective of well-represented behavioral indicators.

**Exploratory Factor Analysis of the 30 Retained Items**

A subsequent EFA was conducted on the 30 items that were retained during the full-scale administration (Table 7). Once items determined to be problematic were removed, the factor analysis was conducted again to establish a more parsimonious factor structure that could be confirmed in a later analysis. Using the same sample ($N = 68$) and only the 30 retained items, another exploratory factor analysis using principal axis factoring (PAF) and oblique (Direct Oblimin) rotation was conducted to determine potential underlying dimensions within the items. This followed the same procedure outlined when all items were analyzed, including the use of Spearman’s rho correlations as the basis for analysis.

The data suggest an adequate relationship among items was achieved with a KMO value of .722, improving upon the .639 obtained in the initial analysis. Bartlett’s test of sphericity was again statistically significant ($\chi^2 = 1194.159, df = 435, p < .001$), indicating the correlation matrix is not an identity matrix and factoring is appropriate to pursue given the data. Factor rotation also followed the Direct Oblimin procedure (i.e., $\delta = 0$) to obtain a meaningful factor solution. The scree plot indicated approximately six factors prior to rotation, while nine factors obtained eigenvalues greater than one. The sums of squared loadings from these components account for 63.55% of the total variance within the data and were evenly distributed following rotation.
Each of the items again demonstrated a moderate relationship with at least one of the factors, with a structure coefficient of .300 or higher. In most cases, items were found to have a structure coefficient value exceeding .300 with more than two factors. Pattern coefficients indicated items were likely to have a unique relationship with a factor while controlling for the correlation among factors, but coefficients greater than .300 were found for multiple factors with two items (i.e., Items 12 and 34). The instances of multiple loadings become problematic in defining the factors and conducting later tests on the structure. Similar to the first analysis, Item 4 is the only item that does not obtain a pattern coefficient of at least .300 with any of the factors. The breakdown of items into factors with pattern coefficients is presented in Table 7.

An examination of the factor correlation matrix indicated that the nine new factors demonstrated little statistical relationship with one another. Factor correlations ranged from .347 to -.331, indicating a very weak relationship between factors. Coefficient Alpha was also calculated for each of the factors’ component items. Obtained Alpha values ranged from .415 (Factor 1) to .735 (Factor 4). The internal consistency was low for Factor 3 (α = .524) due to Item 32’s negative pattern coefficient (-.443), but for most factors exceeded .600. Overall, there are nine underlying factors to which at least two items adequately ascribe, leading to potential in confirming this factor structure with a larger sample.
Table 7.

*Results of the Exploratory Factor Analysis with Items Removed*

<table>
<thead>
<tr>
<th>Item</th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
<th>Factor 4</th>
<th>Factor 5</th>
<th>Factor 6</th>
<th>Factor 7</th>
<th>Factor 8</th>
<th>Factor 9</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.) I acknowledge visitors and welcome them.</td>
<td>0.709</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20.) I focus on patients/visitors when they communicate with me.</td>
<td></td>
<td>0.583</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8.) I am genuine when interacting with patients/visitors.</td>
<td></td>
<td></td>
<td>0.460</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28.) I encourage patients/visitors to ask questions.</td>
<td></td>
<td></td>
<td></td>
<td>0.906</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25.) I am forthcoming in sharing what patients can expect from any procedure.</td>
<td></td>
<td></td>
<td></td>
<td>0.565</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13.) I encourage patients/visitors to discuss or present any concerns throughout their visit.</td>
<td></td>
<td></td>
<td></td>
<td>0.548</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27.) I seek to build a positive, helping relationship with patients/visitors.</td>
<td></td>
<td></td>
<td></td>
<td>0.305</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26.) I am prompt in my service delivery.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.681</td>
<td></td>
<td></td>
<td></td>
<td>0.400</td>
</tr>
<tr>
<td>12.) I can follow through on my responsibilities and still meet patients' expectations.</td>
<td></td>
<td></td>
<td></td>
<td>0.586</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32.) I explain my title or role.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.443</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16.) I clearly communicate to patients/visitors what they can expect from me.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.343</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33.) I greet others warmly.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.619</td>
<td></td>
</tr>
<tr>
<td>17.) I contact the appropriate person to address patients'/visitors' concerns.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.577</td>
<td></td>
<td></td>
</tr>
<tr>
<td>34.) I do my best to assure patients'/visitors' needs are met.</td>
<td></td>
<td></td>
<td></td>
<td>-0.405</td>
<td></td>
<td></td>
<td>0.576</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item</td>
<td>Factor Coefficient</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>--------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9.) I treat patients with care.</td>
<td>0.561</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>39.) I do my best to show compassion for patients or families.</td>
<td>0.392</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40.) I encourage patients/visitors to ask for help when necessary.</td>
<td>0.663</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>44.) I strive to be professional in my interactions.</td>
<td>0.619</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18.) I demonstrate an interest in patients/visitors and their needs.</td>
<td>0.373</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24.) I work collaboratively with my colleagues.</td>
<td>0.664</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>37.) I am sincere in my apologies.</td>
<td>0.603</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>45.) I recognize that each patient/visitor has a unique experience and circumstances.</td>
<td>0.542</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31.) I make eye contact when speaking with patients/visitors.</td>
<td>0.849</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14.) &quot;Polite&quot; and &quot;thoughtful&quot; are two words patients/visitors use to describe me.</td>
<td>0.408</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.) I am knowledgeable about patient care.</td>
<td>0.729</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21.) I have a strong impact on patient experience.</td>
<td>0.634</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.) I introduce myself when I meet patients/visitors.</td>
<td>0.258</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>46.) I close every interaction by saying &quot;Thank you&quot; to patients/visitors.</td>
<td>0.261</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.) I use patients'/visitors' preferred name when I speak with them.</td>
<td>0.261</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>41.) When I forward a patient's/visitor's concern, I follow up to ensure it has been handled.</td>
<td>0.687</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note. Coefficients less than the absolute value of .400 are suppressed unless no factor coefficient exceeds .400 for an item.*
Vignettes

The number of participants attempting each of the vignettes ranged from 76 to 81, with the last vignette most likely to be left incomplete (Table 8). The responses to the vignettes were scored for accuracy, with participants receiving scores ranging from one to eight. Participants who did not answer any of these items were the only respondents to record a score of zero, meaning that all employees provided at least one correct answer when they attempted the vignettes. When those missing all responses were excluded, the mean vignette score for the sample was 4.53 ($SD = 1.51$), with only two participants recording a perfect score (2.40%). Table 8 presents the item-level results of the vignettes.

Table 8

<table>
<thead>
<tr>
<th>Vignette</th>
<th>Percent Correct</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>40.96</td>
<td>81</td>
</tr>
<tr>
<td>2</td>
<td>90.24</td>
<td>81</td>
</tr>
<tr>
<td>3</td>
<td>50.00</td>
<td>81</td>
</tr>
<tr>
<td>4</td>
<td>30.12</td>
<td>81</td>
</tr>
<tr>
<td>5</td>
<td>62.81</td>
<td>78</td>
</tr>
<tr>
<td>6</td>
<td>58.97</td>
<td>78</td>
</tr>
<tr>
<td>7</td>
<td>82.71</td>
<td>81</td>
</tr>
<tr>
<td>8</td>
<td>46.05</td>
<td>76</td>
</tr>
</tbody>
</table>

Based on the varying response rates, participants are believed to have skipped a vignette if they were unsure of their response. Upon closer examination of the database and response strings, respondents appear to have chosen which items to which they wanted to respond, even skipping some rating-scale items to test their knowledge on the vignette portion of the instrument. Altogether, only two vignettes (i.e., 2 and 7) were met
with over 80% of responses correct. These items in particular will be considered for revision in the full-scale instrument.

Vignette scores were also compared based on the nature of employees’ current roles. Total vignette scores were calculated for those with \( n = 57 \) and without \( n = 24 \) direct patient care contact. The mean scores of participants who report direct patient care contact \( (M = 4.56, SD = 1.50) \) and those who do not \( (M = 4.46, SD = 1.59) \) were compared using an independent sample t-test, which suggests there was no significant difference between the two groups \( (t = .278, df = 79, p > .05) \). Ultimately, employees who have more visitor-oriented interactions are equally as able as those with primarily patient-oriented interaction to be aware of the expectations and to select their behaviors accordingly.

**Finalizing the Full-Scale Instrument**

Following the refinement of rating scale items and vignettes, the full-scale instrument consists of 47 total items: the question of direct patient care contact, 39 self-report behavioral items, and seven vignettes. Within the 39 items for which employees were asked to provide ratings on the frequency of their behavior, nine items are queried from their colleagues’ point of view while the remaining 30 assume the perspective of patients/visitors.

When the vignettes were considered, two warranted further investigation based on correct response rates higher than 80%. One of the scenarios (Vignette 7) was slightly re-worded along with the response options with the intention of making it more difficult for the full-scale administration. The other scenario (Vignette 2) was not as easily modified,
so it was dropped entirely from the final instrument. Response choices were also randomly shuffled to remove any potential concerns with ordering. There were no other modifications recommended based on the results, leading the full-scale instrument to consist of 7 vignettes.

The seven vignettes are presented between the patient/visitor and colleague rating-scale items, in that order. The vignettes were inserted between rating-scale items to signal a differentiation in focus from patients and visitors to colleagues through provision of a different question format. Altogether, the 101 responses requested on the pilot instrument were reduced to 47 on the full-scale revision, which was expected to significantly reduce completion time while providing valuable data. The final version of the full-scale instrument is provided in Appendix B.

**Part II: Full-Scale Administration**

The results presented in the full-scale section include another item analysis, principal components and confirmatory factor analyses, and Rasch model investigations into item difficulty and fit, response category usage, dimensionality, and DIF.

**Sample**

The initial request for participation in the full-scale administration phase was sent via company e-mail to 1,255 employees who participated in the training. To garner a larger sample, employees participating in the program for the reasons of service recovery \( (n = 518) \) were added to those newly-hired \( (n = 737) \) to constitute the total sample for full-scale administration. There were no modifications to the training under either
circumstance, so they were combined for the sample. Reminders to non-responders were sent weekly from the institution’s Office of Patient Experience for the 12 weeks the instrument was open for data collection.

Of the original sample of 1,255 employees, 452 provided consent and began the pilot instrument (36.01%). The number of completed instruments was 266, comprising 21.19% of the entire sample. The rate of completion was much improved over the pilot testing phase, likely due to the shorter instrument. The response rate in excess of 20% is much more in line with the expected rate using the online survey method (Nulty, 2008). While the rate of completed surveys was better for the full-scale administration, it remained less than the predicted sample size and is a cause for caution in the analysis and interpretation of the results. The instrument was developed for use with a targeted and captive sample that could not be expanded given the scope of the research. The final sample used in the analysis consisted mostly of females (83.33%) and many reported a position involving direct patient care contact (67.05%). A breakdown of frequencies for each group is presented below in Table 9.

Table 9.

*Cross-Tabulation between Gender and Direct Patient Care Contact on the Full-Scale Administration*

<table>
<thead>
<tr>
<th>Gender</th>
<th>Direct Care Contact (%)</th>
<th>No Direct Care Contact (%)</th>
<th>Total (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>155 (60.07)</td>
<td>60 (23.25)</td>
<td>215 (83.33)</td>
</tr>
<tr>
<td>Male</td>
<td>18 (6.97)</td>
<td>25 (9.68)</td>
<td>43 (16.66)</td>
</tr>
<tr>
<td>Total</td>
<td>173 (67.05)</td>
<td>85 (32.94)</td>
<td>258* (100.00)</td>
</tr>
</tbody>
</table>

*Note. The table only presents respondents who identified both gender and direct patient care contact. Those not reporting a gender were excluded, so the value is different from the total sample of 266 in the analysis.*
Item Analysis of 30 “Patient/Visitor” Items

The 30 rating-scale items assessing patient/visitor behavior provided evidence of strong internal consistency, with a Coefficient Alpha of .920 for the entire scale. Thus, the internal consistency was enhanced through the iterative process of pilot testing and removing items that were considered troublesome within the measure. Item means remained exceptionally high, ranging from 4.43 (i.e., Item 6: “I am forthcoming in sharing what patients can expect from any measure.”) to 4.93 (i.e., Item 6: “I treat patients with care.”). This is of concern given the response scale ranging from one to five, suggesting the presence of a ceiling effect. Further, item standard deviations fell below .50 for 19 of the items, indicating the data represent a skew and leptokurtic distribution. Based on the item distributions, employees consistently report that they engage in these behaviors frequently when interacting with patients, visitors, and their colleagues.

Item-total correlations ranged from .423 (i.e., Item 5: “I am genuine when interacting with patients/visitors.”) to .693 (i.e., Item 12: “I demonstrate an interest in patients/visitors and their needs.”), indicating that all the items adequately represent the construct. Similarly, review of the inter-item correlation matrix provided no concerns as the inter-item correlations were all positive and many fell within the moderate range (i.e., .25 < ρ < .50). Lastly, there were no threats to internal consistency found upon examination of Alpha if Item Removed values (Table 10). All values were calculated to be lower than the obtained Coefficient Alpha (α = .920) for the scale. Item mean, standard deviation, and Alpha if Item Removed statistics are presented in Table 10.
Table 10.

*Descriptive Statistics and Alpha if Item Removed for Patient/Visitor Items on the Full-Scale Instrument*

<table>
<thead>
<tr>
<th>Patient/Visitor Rating Item</th>
<th>Mean</th>
<th>SD</th>
<th>Alpha if Removed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.) I am knowledgeable about patient care.</td>
<td>4.44</td>
<td>0.82</td>
<td>0.919</td>
</tr>
<tr>
<td>2.) I acknowledge visitors and welcome them.</td>
<td>4.78</td>
<td>0.60</td>
<td>0.916</td>
</tr>
<tr>
<td>3.) I introduce myself when I meet patients/visitors.</td>
<td>4.69</td>
<td>0.70</td>
<td>0.918</td>
</tr>
<tr>
<td>4.) I use patients'/visitors' preferred name when I speak with them.</td>
<td>4.67</td>
<td>0.65</td>
<td>0.918</td>
</tr>
<tr>
<td>5.) I am genuine when interacting with patients/visitors.</td>
<td>4.91</td>
<td>0.37</td>
<td>0.918</td>
</tr>
<tr>
<td>6.) I treat patients with care.</td>
<td>4.93</td>
<td>0.35</td>
<td>0.918</td>
</tr>
<tr>
<td>7.) I contact the appropriate person to address patients'/visitors' concerns.</td>
<td>4.79</td>
<td>0.43</td>
<td>0.917</td>
</tr>
<tr>
<td>8.) I can follow through on my responsibilities and still meet patients' expectations.</td>
<td>4.76</td>
<td>0.47</td>
<td>0.916</td>
</tr>
<tr>
<td>9.) I encourage patients/visitors to discuss or present any concerns throughout their visit.</td>
<td>4.67</td>
<td>0.63</td>
<td>0.916</td>
</tr>
<tr>
<td>10.) &quot;Polite&quot; and &quot;thoughtful&quot; are two words patients/visitors use to describe me.</td>
<td>4.81</td>
<td>0.44</td>
<td>0.916</td>
</tr>
<tr>
<td>11.) I clearly communicate to patients/visitors what they can expect from me.</td>
<td>4.73</td>
<td>0.49</td>
<td>0.915</td>
</tr>
<tr>
<td>12.) I demonstrate an interest in patients/visitors and their needs.</td>
<td>4.84</td>
<td>0.39</td>
<td>0.915</td>
</tr>
<tr>
<td>13.) I focus on patients/visitors when they communicate with me.</td>
<td>4.90</td>
<td>0.31</td>
<td>0.917</td>
</tr>
<tr>
<td>14.) I have a strong impact on patient experience.</td>
<td>4.56</td>
<td>0.78</td>
<td>0.919</td>
</tr>
<tr>
<td>15.) I am prompt in my service delivery.</td>
<td>4.77</td>
<td>0.46</td>
<td>0.917</td>
</tr>
<tr>
<td>16.) I work collaboratively with my colleagues.</td>
<td>4.89</td>
<td>0.33</td>
<td>0.918</td>
</tr>
<tr>
<td>17.) I am forthcoming in sharing what patients can expect from any procedure.</td>
<td>4.43</td>
<td>1.01</td>
<td>0.919</td>
</tr>
<tr>
<td>18.) I seek to build a positive, helping relationship with patients/visitors.</td>
<td>4.85</td>
<td>0.45</td>
<td>0.916</td>
</tr>
<tr>
<td>19.) I make eye contact when speaking with patients/visitors.</td>
<td>4.84</td>
<td>0.56</td>
<td>0.918</td>
</tr>
<tr>
<td>20.) I explain my title or role.</td>
<td>4.72</td>
<td>0.62</td>
<td>0.918</td>
</tr>
<tr>
<td>21.) I greet others warmly.</td>
<td>4.93</td>
<td>0.29</td>
<td>0.916</td>
</tr>
<tr>
<td>22.) I do my best to assure patients'/visitors' needs are met.</td>
<td>4.89</td>
<td>0.33</td>
<td>0.916</td>
</tr>
</tbody>
</table>
Table 10. (continued)

<table>
<thead>
<tr>
<th>Patient/Visitor Rating Item</th>
<th>Mean</th>
<th>SD</th>
<th>Alpha if Removed</th>
</tr>
</thead>
<tbody>
<tr>
<td>23.) I encourage patients/visitors to ask questions.</td>
<td>4.78</td>
<td>0.48</td>
<td>0.916</td>
</tr>
<tr>
<td>24.) I am sincere in my apologies.</td>
<td>4.91</td>
<td>0.31</td>
<td>0.916</td>
</tr>
<tr>
<td>25.) I do my best to show compassion for patients or families.</td>
<td>4.91</td>
<td>0.32</td>
<td>0.916</td>
</tr>
<tr>
<td>26.) When I forward a patient's/visitor's concern, I follow up to ensure it has been handled.</td>
<td>4.48</td>
<td>0.73</td>
<td>0.918</td>
</tr>
<tr>
<td>27.) I offer help and encourage patients/visitors to ask for help when necessary</td>
<td>4.81</td>
<td>0.44</td>
<td>0.915</td>
</tr>
<tr>
<td>28.) I strive to be professional in my interactions.</td>
<td>4.90</td>
<td>0.32</td>
<td>0.917</td>
</tr>
<tr>
<td>29.) I recognize that each patient/visitor has a unique experience and circumstances.</td>
<td>4.91</td>
<td>0.31</td>
<td>0.917</td>
</tr>
<tr>
<td>30.) I close every interaction by saying &quot;Thank you&quot; to patients/visitors.</td>
<td>4.69</td>
<td>0.62</td>
<td>0.918</td>
</tr>
</tbody>
</table>

*Note. N = 266 for all.*

**Item Analysis of 9 “Colleague” Items**

The nine rating-scale items assessing colleague behavior provided evidence of moderate internal consistency, with a Coefficient Alpha of .607 for the scale. This was expected, as the internal consistency was strong ($\alpha = .864$) when a troublesome item (i.e., Item 4: “My colleagues would say that…If I’m having a challenging day, patients/visitors can tell.”) was removed. The descriptive statistics and Alpha if removed for the colleague items are presented in Table 11. Item 4 obtained an item mean of 2.68, suggesting this behavior is exhibited between “rarely” and “sometimes”. The remaining item means remained exceptionally high, ranging from 4.72 (i.e., Item 7: “I anticipate the needs of patients/visitors.”) to 4.94 (i.e., Item 2: “I am considerate to patients/visitors.”) and Item 9: “I am mindful of patients’ privacy and dignity.”). Item standard deviations fell below .50 for six of the nine items, indicating the data represent a skew and leptokurtic distribution.
Based on the item distributions, employees consistently report that their colleagues would rate the frequency of these behaviors as occurring almost always.

Table 11.

Descriptive Statistics and Alpha if Item Removed for Colleague Items on the Full-Scale Instrument

<table>
<thead>
<tr>
<th>Colleague Rating Item</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Alpha if Removed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.) I treat patients with care.</td>
<td>249</td>
<td>4.92</td>
<td>0.28</td>
<td>0.561</td>
</tr>
<tr>
<td>2.) I am considerate to patients/visitors.</td>
<td>252</td>
<td>4.94</td>
<td>0.25</td>
<td>0.561</td>
</tr>
<tr>
<td>3.) I focus on patients/visitors when they communicate with me.</td>
<td>252</td>
<td>4.93</td>
<td>0.27</td>
<td>0.561</td>
</tr>
<tr>
<td>4.) If I’m having a challenging day, patients/visitors can tell.</td>
<td>253</td>
<td>2.68</td>
<td>1.44</td>
<td>0.864</td>
</tr>
<tr>
<td>5.) I am prompt in my service delivery.</td>
<td>251</td>
<td>4.76</td>
<td>0.44</td>
<td>0.542</td>
</tr>
<tr>
<td>6.) Each interaction I have with patients is authentic and individualized.</td>
<td>253</td>
<td>4.81</td>
<td>0.50</td>
<td>0.537</td>
</tr>
<tr>
<td>7.) I anticipate the needs of my patients/visitors.</td>
<td>252</td>
<td>4.72</td>
<td>0.52</td>
<td>0.531</td>
</tr>
<tr>
<td>8.) I strive to be professional in my interactions.</td>
<td>252</td>
<td>4.92</td>
<td>0.29</td>
<td>0.552</td>
</tr>
<tr>
<td>9.) I am mindful of patients’ privacy and dignity.</td>
<td>249</td>
<td>4.94</td>
<td>0.26</td>
<td>0.561</td>
</tr>
</tbody>
</table>

Note. Coefficient Alpha for all items = .607.

Several items presented in the colleague section were redundant from those previously administered in the patient/visitor rating section. Correlations between sections ranged from .391 (Item 3: “I focus on patients/visitors when they communicate with me.”) to .652 (Item 8: “I strive to be professional in my interactions.”), suggesting consistency between colleague and patient/visitor perspectives on the self-report. Five items were included within this section (i.e., Items 2, 4, 6, 7, and 9) based on reasonable factor loadings in the pilot sample EFA. Due to the fewer number of responses, limited variability evidenced in the item analysis, and the number of patient/visitor items
representing each service behavior, these nine items were no longer included in further analyses.

**Vignettes**

The number of participants attempting each of the vignettes ranged from 284 to 300, with the fifth vignette most likely to be left incomplete. Some vignettes were challenging for respondents, as four of them (i.e., Vignettes 1, 2, 3, and 6) were met with less than 50% correct responses. Table 12 presents the item-level results of the vignettes.

Table 12.

<table>
<thead>
<tr>
<th>Vignette</th>
<th>Percent Correct</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>36.67</td>
<td>300</td>
</tr>
<tr>
<td>2</td>
<td>45.92</td>
<td>294</td>
</tr>
<tr>
<td>3</td>
<td>22.78</td>
<td>294</td>
</tr>
<tr>
<td>4</td>
<td>68.40</td>
<td>288</td>
</tr>
<tr>
<td>5</td>
<td>67.25</td>
<td>284</td>
</tr>
<tr>
<td>6</td>
<td>48.07</td>
<td>285</td>
</tr>
<tr>
<td>7</td>
<td>74.31</td>
<td>288</td>
</tr>
</tbody>
</table>

For individuals, the responses to the vignettes were scored for accuracy, with participants receiving scores ranging from zero to eight. When those missing all responses were excluded, a total sample of 273 respondents provided answers to all seven vignettes. Only one participant attempted all seven vignettes but did not answer any correctly. The mean vignette score for the sample was 3.63 ($SD = 1.34$), with only three participants recording a perfect score (1.09%). Compared with the pilot testing results, the changes made to the vignettes yielded items that were more difficult than the previous administration.
Two comparisons were made using the total vignette score to determine if there were differences within the sample. The first comparison was made to see if there was a difference between genders on the vignette total score. Though there were many more females \((n = 203)\) than males \((n = 43)\), there was no significant difference between genders \((t = 1.939, df = 244, p > .05)\). The male mean \((4.05, SD = 1.17)\) was higher than that of females \((3.62, SD = 1.34)\), but the difference was not significant. The second comparison sought to determine if direct patient care contact had an impact on vignette score. More respondents reported direct patient care contact \((n = 183)\) than those that did not \((n = 90)\) when examining participants with completed vignettes. Means were very close, while those without direct patient care contact \((3.69, SD = 1.35)\) scored higher than those with direct patient care contact \((3.60, SD = 1.34)\), but that difference was not significant \((t = 0.507, df = 271, p > .05)\). Ultimately, there was no difference between the genders or respondents with and without direct patient care contact on their vignette score.

**Data Reduction**

Following the item analysis, the initial research plan was to confirm the factor structure obtained through the pilot testing phase using the full-scale sample. Due to the questionable validity of the exploratory factor analysis based on the small sample size obtained in the pilot phase, as well as the strong inter-relationships among items, an additional step of data reduction and model simplification took place. The number of variables was reduced prior to the confirmatory analyses to alleviate potential concerns with multicollinearity and ensure a fundamental, logical structure could be tested.
First, the 68 responses completed during the pilot testing phase were added to the full-scale administration sample of 266 to generate a total sample of 334 respondents. Combining pilot and full-scale administration samples was done to increase the size of the overall sample, in hopes of providing more valid results in the confirmatory factor and Rasch analyses. This was feasible since the items retained for the full-scale administration were identical to those presented in the pilot testing phase. Further, all responses were collected by employees who had participated in the training during the allotted time, so there are no threats to validity based on time since training. Ideally, the full-scale administration sample would have been large enough the pilot phase data would have remained independent. However, combining respondents from both instruments was done to enhance the accuracy and reliability of the following analyses.

**Principal Components Analysis**

A subsample of 150 respondents was randomly selected from the combined sample of 334 and analyzed using Principal Components Analysis (PCA). The current sample of 150 was drawn randomly to satisfy the condition of a minimum sample size to variable ratio of five to one (Gorsuch, 1983), given the 30 items presented on both instruments.

In this analysis, PCA was utilized to serve the purpose of reducing items into components with as little a loss of information as possible (Thompson, 2004). It was anticipated that a few of the 30 items may be identified as problematic, but the results would serve to inform the reduction of items into fewer, more manageable components.
Because PCA is based on the assumption of perfect reliability of the data (Fabrigar, Wegener, MacCallum, & Strahan, 1999), the analysis was employed to assume and account for all of the variance among the items (Thompson, 2004). Due to the already limited variability in the responses based on the high means and low standard deviations, PCA can be argued to be a more useful analysis method by including all possible variability among the responses, thus more helpful in determining the most useful items to be used in further analyses.

PCA was used in this instance instead of traditional EFA for several reasons. PCA is different from traditional EFA, though PCA is often confused with and misused as a substitute or variant of EFA (Henson & Roberts, 2006). Research suggests that in some cases, PCA serves as a first step of item screening, prior to subjecting items to EFA (Matsunaga, 2010). Though it was not planned to be included at this stage of the analysis, PCA was seen as a necessary first step to reduce the items into fewer, more manageable components using all available variance within the items. Conducting a PCA is often used to summarize the information available provided by a given set of variables (i.e., items) and reduce it into a fewer number of components (Fabrigar et al., 1999). The intent of employing PCA was to use the subsample to determine components through item reduction, which would inform the confirmatory analysis used later. Further, EFA was not used in this instance because of the limited common variance available among the item responses. In order to make use of as much variability within the data as possible, PCA accounts for all variability where EFA does not and, therefore, was chosen here to
be the more appropriate analysis. Ultimately, the results of the PCA were expected to provide a framework of components that could be validated using CFA.

Several items were determined to be problematic and removed prior to conducting the PCA. Seven items with exceptionally high means (i.e., greater than 4.90 on a 5-point scale) and four items included to evaluate global behaviors of service quality were not helpful in defining the components, so they were not included in the PCA. Items with very high means were removed because of the lack of variability within them. The global items were designed and added in the development phase to serve as validation of the other items and evaluate if they ascribed to one specific behavior over another, to which they did not. These global behaviors were omitted in the analysis because they were intended to reflect a combination of the individual behaviors portrayed in the other items. These global items were retained after the pilot study to provide support for construct validity when examining the correlations between these all-encompassing items and other individual behavioral indicators. Including them in an analysis of components could be problematic because these all-encompassing items are expected to align with each of the components and are not theoretically aligned with one specific component. Based on these criteria, 11 items were removed and the analysis proceeded with the remaining 19 items.

The PCA was conducted using the subsample of 150 respondents and their responses to the 19 items. The original component solution was rotated using Varimax rotation, an orthogonal rotation that provides components that are uncorrelated with one another (Pett, et al., 2003). This rotation method was chosen as a means of maximizing
the differences between components with the intention of alleviating looming issues of multicollinearity in the CFA. While there are concerns with utilizing orthogonal rotations in the social science literature (see Pett et al., 2003), a solution providing strongly related components would be counter-productive in the current investigation. The overarching purpose of the PCA was to determine underlying components of the variables that are statistically unrelated to one another.

The data suggest a moderate relationship among items was achieved with a KMO value of .815, improving upon the .722 obtained in the pilot phase with 30 items. Bartlett’s test of sphericity was again statistically significant ($\chi^2_B = 1307.749$, $df = 171$, $p < .001$), indicating the correlation matrix is not an identity matrix and PCA is appropriate to pursue given the data. The scree plot indicated approximately five components, verified by obtained eigenvalues greater than one. Following rotation, the sums of squared loadings from these components account for 64.68% of the total variance within the data and were evenly distributed following the orthogonal rotation, with all components accounting for at least 7.50% of the variance within items.

Each of the items was found to demonstrate a relationship with at least one of the components (Table 13). A rotated component coefficient of .450 or higher was achieved for all of the 19 items. In a few cases, items were found to have loading coefficients exceeding .400 with more than two components. The five cases of multiple loadings may become problematic in defining the components and conducting later tests on the structure, but the situations of dual loadings are handled considering each item individually in the following analyses. For the discussion of internal consistency within
the components, the items with dual loadings were included where they obtained the highest loading coefficient. For instance, Item 3 obtained loading coefficients of .771 and .421 on components 2 and 5, respectively. For the analysis of internal consistency, Item 3 was included within component 2. The breakdown of items into components with rotated loading coefficients is presented in Table 13.
Table 13.

**Results of the Principal Components Analysis with Varimax Rotation**

<table>
<thead>
<tr>
<th>Item</th>
<th>Component 1</th>
<th>Component 2</th>
<th>Component 3</th>
<th>Component 4</th>
<th>Component 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>13.) I focus on patients/visitors when they communicate with me.</td>
<td>.761</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12.) I demonstrate an interest in patients/visitors and their needs.</td>
<td>.734</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10.) “Polite” and “thoughtful” are two words patients/visitors would use to describe me.</td>
<td>.716</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27.) I offer help and encourage patients/visitors to ask for help when necessary</td>
<td>.695</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.) I contact the appropriate person to address patients'/visitors' concerns.</td>
<td>.618</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18.) I seek to build a positive, helping relationship with patients/visitors.</td>
<td>.565</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30.) I close every interaction by saying &quot;Thank you&quot; to patients/visitors.</td>
<td>.476</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.) I use patients'/visitors' preferred name when I speak with them.</td>
<td>.800</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.) I introduce myself when I meet patients/visitors.</td>
<td>.771</td>
<td>.421</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.) I acknowledge visitors and welcome them.</td>
<td>.581</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26.) When I forward a patient's/visitor's concern, I follow up to ensure it has been handled.</td>
<td>.511 .463</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19.) I make eye contact when speaking with patients/visitors.</td>
<td>.750</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15.) I am prompt in my service delivery.</td>
<td>.697</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16.) I work collaboratively with my colleagues.</td>
<td>.508</td>
<td>.584</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11.) I clearly communicate to patients/visitors what they can expect from me.</td>
<td>.491 .423</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9.) I encourage patients/visitors to discuss or present any concerns throughout their visit.</td>
<td>.785</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23.) I encourage patients/visitors to ask questions.</td>
<td>.765</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17.) I am forthcoming in sharing what patients can expect from any procedure.</td>
<td>.492 .628</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20.) I explain my title or role.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.735</td>
</tr>
</tbody>
</table>
The first component, consisting of seven items (i.e., Items 7, 10, 12, 13, 18, 27, and 30), seems to reflect empathy or active listening, two of the nine Expected Service Behaviors from which the items were developed. The next component consists of three items (i.e., Items 2, 3, and 4) which examine acknowledgement and welcoming of patients and visitors, another of the expected behaviors. Also included in this component was Item 26 (i.e., “When I forward a patient’s/visitor’s concern, I follow up to ensure it has been handled”), which also exhibited a loading on Component three. Component three consists of items related to professionalism or professional duties, as it included Items 11, 15, 16, and 19. The fourth component accounted for Items 9, 17, and 23, and was representative of behaviors indicating communication between employees and patients or visitors. The components and items were found to align more closely with the behaviors from which they were developed than in previous analyses, suggesting the PCA was useful in establishing unique components given the items. This component solution was used as the basis for the confirmatory analysis that follows.

Coefficient Alpha was calculated for each of the components’ items to evaluate internal consistency and assist in identifying problematic items. Obtained Alpha values ranged from .665 (Component 4) to .832 (Component 1), which suggest at least moderate internal consistency of the items within components. Alpha was not calculated for Component 5, since it consisted of only one item (i.e., Item 20), nor was it included in the confirmatory analysis due to the likelihood it would contribute to an improper solution (Brown, 2015). Coefficient Alpha for Component 2 ($\alpha = .721$), would have increased with removal of Item 26, which did not align in content with the other items and also
loaded on Component 3. Component 4 obtained a Coefficient Alpha of .724, where Item 11 more closely aligned based on loading values. Coefficient Alpha for Component 4 ($\alpha = .664$) could have been improved with removal of Item 17, another item that loaded on more than one component. Some of the items with dual loadings were problematic when evaluated according to the highest loading coefficient (i.e., Items 26 and 17), leading to weaker internal consistency. Overall, there are four underlying components to which at least two items adequately ascribe, leading to potential in confirming this factor structure with the full sample.

**Confirmatory Factor Analysis**

The confirmatory analysis was completed using LISREL 9.1 (Joreskog & Sorbom, 2013). The CFA was conducted using data from all 334 participants to validate the relationship between the items and components found as the results of the PCA. Thus, the purpose of the CFA was to validate the structure of four components with 18 items. Based on the notion that having multiple indicators per factor decreases the likelihood of improper solutions (Brown, 2015), the fifth component was excluded from this analysis because it comprised only one item. As defined earlier, Component 1 was made up of 7 items, while Components 2 and 3 each consisted of four items, and Component 4 included three items. The dual loadings identified in the PCA were accounted for by including each item within the component to which it achieved higher loadings (e.g., Item 3 was included in Component 2 with a loading of .511, instead of Component 3 with a loading of .463). The initial model is depicted in Figure 1.
Figure 1. The initial model. This figure represents the item and component structure based on the results of the PCA.

The confirmatory analyses utilized the Weighted Least Squares (WLS) estimation method due to the nature of the non-normal, ordinal data. The WLS method assumes no distribution, so it is used in cases of non-normality (Schumacker & Lomax, 2010). This was chosen over the default Maximum Likelihood (ML) estimation method because it provides more accurate parameter estimates and the ML method tends to inflate some statistics when used with non-normal data (Mindrila, 2010). The asymptotic covariance matrix was also used as a weight matrix to correct for bias in standard errors and fit statistics. These modifications to the default methods were made to account for the
skewed distribution of responses within the items and were expected to provide more unbiased fit statistics and parameter estimates.

The results of the initial CFA provided a poorly-fitting model and concerns with multicollinearity. All paths were found to be significant \( p < .05 \). However, the fit indices suggest the data do not fit the model. When the data adequately fit the model, the Chi-square value is non-significant (i.e., greater than .05) and the values for the root mean square error of approximation (RMSEA) and the standardized root mean residual (SRMR) are below .05, but at least less than .08 (Schumacker & Lomax, 2010). The current model found a large, significant Chi-square \( \chi^2 = 1967.227, df = 129, p < .001 \) and an RMSEA well above the cutoff at .207. The standardized root mean residual (SRMR) of .071 could be considered acceptable, but does not indicate model fit when it is the only index that suggests adequate fit. Other fit indices agreed that the data and model were misaligned, with the Goodness of Fit Index (GFI) and the Adjusted Goodness of Fit Index (AGFI) achieving values of .761 and .684, respectively. In well-fitting models, these values are at least .90 (Schumacker & Lomax, 2010). Taken together, the model proposed by the PCA required further considerations and modifications to be upheld using CFA techniques. Table 14 presents the model fit indices from the CFA and the acceptable range for each.
Table 14.

*Initial Confirmatory Factor Analysis Results: Fit Indices*

<table>
<thead>
<tr>
<th>Index</th>
<th>Acceptable Range for Model Fit (Schumacker &amp; Lomax, 2010)</th>
<th>Obtained Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-Square</td>
<td>( p &gt; .001 )</td>
<td>( p &lt; .001 )</td>
</tr>
<tr>
<td>Root Mean Square Error of Approximation (RMSEA)</td>
<td>Less than .05</td>
<td>.207</td>
</tr>
<tr>
<td>Standardized Root Mean Residual (SRMR)</td>
<td>Less than .05</td>
<td>.071</td>
</tr>
<tr>
<td>Goodness of Fit Index (GFI)</td>
<td>Greater than .90</td>
<td>.761</td>
</tr>
<tr>
<td>Adjusted Goodness of Fit Index (AGFI)</td>
<td>Greater than .90</td>
<td>.684</td>
</tr>
</tbody>
</table>

The first issue presented in the data is multicollinearity, which is the presence of two or more variables that are highly inter-related when used for prediction (Stevens, 2009). Strong inter-item correlations are problematic in CFA and are evidence of redundancy of items on the measurement (Schumacker & Lomax, 2010). Essentially, at least two variables included in the analysis overlap and explain the same variance within factors, which is a model specification error. The presence of items with high inter-correlations (i.e., greater than .50) is problematic because the model is not parsimonious.

To develop a more parsimonious model, the results of the PCA were revisited to select the highest-loading items within each component. Component 1’s four highest-loading items were retained (i.e., Items 10, 12, 13, and 27), while the three highest-loading items were retained from the remaining components (i.e., Items 2, 3, and 4 for Component 2, Items 15, 16, and 19 for Component 3, and Items 9, 23, and 17 for Component 4). The five items that were removed were mis-fitting within their
components compared with those that were retained and two of them (i.e., Items 26 and 11) had dual-loadings, which is believed to have caused measurement errors.

The more parsimonious model consisted of 13 items comprising four factors. As was the case previously, the data did not demonstrate adequate fit with the model and multicollinearity was identified. The Chi-square remained significant ($\chi^2 = 416.723, df = 59, p < .001$), but was substantially lower than the previous model. The RMSEA (.135) suggested poor fit between the data and the model, while the SRMR (.054) suggested moderate fit. The goodness of fit indices were higher, but still not high enough to support the case of good fit, with a GFI of .853 and an AGFI of .773. The model fit indices are presented for the modified and preliminary models in Table 15. All paths were significant (i.e., $p < .05$), but the results reflect continued concerns with overall model fit.

Examination of the modification indices provided little insight into potential solutions that would assist in further specifying the model.

Table 15.

<table>
<thead>
<tr>
<th>Modified Model Confirmatory Factor Analysis Results: Fit Indices</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Index</strong></td>
</tr>
<tr>
<td>-----------</td>
</tr>
<tr>
<td>Chi-Square</td>
</tr>
<tr>
<td>Root Mean Square Error of Approximation (RMSEA)</td>
</tr>
<tr>
<td>Standardized Root Mean Residual (SRMR)</td>
</tr>
<tr>
<td>Goodness of Fit Index (GFI)</td>
</tr>
<tr>
<td>Adjusted Goodness of Fit Index (AGFI)</td>
</tr>
</tbody>
</table>

*Note.* The preliminary model consists of 18 items within 4 components. The modified model consists of 13 items within 4 components.
To provide further insight into model concerns, the standardized residual values were examined. The residuals represent the amount of misfit between the model and the data. The residuals are standardized, so values greater than the absolute value of two are worth further examination. The largest standardized residual (2.96) was found between Items 19 and 10, while the smallest standardized residual (-2.00) was found between Items 17 and 10. Item 17 was a candidate for removal based on dual-loadings in the rotated component matrix of the PCA. Based on the large standardized residual values, Items 17 and 19 were omitted from subsequent analyses.

Another, more parsimonious confirmatory factor model was developed, using 11 items attributed to four factors. Factor 1 consisted of five items, Factor 2 consisted of three items, and the remaining two factors each consisted of two items apiece. The issues of multicollinearity were rectified in the updated model, so the items which were removed were identified as redundant with those retained. All paths were significant \((p < .05)\) and model fit statistics substantially improved, but still do not indicate adequate model fit. The Chi-square remained significant \((\chi^2 = 239.534, df = 38, p < .001)\) and RMSEA (.126) was still above that of a well-fitting model, but the values for SRMR (.043) and the GFI (.899) are much more representative of better fit. The AGFI improved to .824, but is still below the adequate range. While the data continues to demonstrate poor fit with the model, the fit indices are much more in line with those expected of well-fitting models. Table 16 presents the results of the third model alongside the previous models.
Table 16.

<table>
<thead>
<tr>
<th>Index</th>
<th>Acceptable Range*</th>
<th>Preliminary Model</th>
<th>Modified Model</th>
<th>Third Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-Square</td>
<td>$p &gt; .001$</td>
<td>$p &lt; .001$</td>
<td>$p &lt; .001$</td>
<td>$p &lt; .001$</td>
</tr>
<tr>
<td>Root Mean Square Error of Approximation (RMSEA)</td>
<td>Less than .05</td>
<td>.207</td>
<td>.135</td>
<td>.126</td>
</tr>
<tr>
<td>Standardized Root Mean Residual (SRMR)</td>
<td>Less than .05</td>
<td>.071</td>
<td>.054</td>
<td>.042</td>
</tr>
<tr>
<td>Goodness of Fit Index (GFI)</td>
<td>Greater than .90</td>
<td>.761</td>
<td>.853</td>
<td>.899</td>
</tr>
<tr>
<td>Adjusted Goodness of Fit Index (AGFI)</td>
<td>Greater than .90</td>
<td>.684</td>
<td>.773</td>
<td>.824</td>
</tr>
</tbody>
</table>

*The Acceptable Range is based on Schumacker & Lomax (2010).

The results of the third model are approaching significance, suggesting the model is more clearly and appropriately specified compared with earlier models. Modifications were made to the current model to generate the final model. It is important to note that modifications to the model such as those that follow introduce a shift from a pure confirmatory approach to an approach that is quasi-confirmatory in nature (Stevens, 2009). The third model presented a plausible model that could demonstrate fit given certain modifications.

The modifications made to the model were the inclusion of error covariance terms between several items, which represent the relationship between error terms present in the items (Schumacker & Lomax, 2010). Due to the inter-relatedness of the items evidenced in the item-total and inter-item correlations, it was expected a priori that there would be correlated error terms that assist in generating model fit. Theoretically, models can allow for correlations among indicators collected via a single method (Kenney, 1976;
Marsh, 1989). This process of modifying the model by including error covariance terms has been met with disagreement between researchers (e.g., Saris & Aalberts, 2003). All things considered, the process of adding error covariance terms was beneficial in improving model fit given the method of data collection and the relationship among the items.

The inclusion of error covariance terms followed an iterative process, adding the term most influential to model fit given the previous model. Those changes that were used formatively led to enhanced model fit statistics and did not substantially change the model tested. Throughout the testing of modifications, all paths remained significant ($p < .001$) and no factor loadings changed from the third model, presented previously. One error variance, that for Item 9, was not significant ($p > .05$), while all other items included significant error variances ($p < .05$). There were a total of eight error covariances included in the final model, only one of which was significant ($p < .05$). The final confirmatory factor model is presented in Figure 2, with dashed gray lines representing the non-significant error variances (i.e., Item 9) and covariances.
Figure 2. The final model. This figure represents the item and component structure of the model best fit to the results of the PCA. Dashed, gray paths are not significant.

The inclusion of multiple error covariances provides evidence of possible model mis-specification. The addition of correlated error terms, even those that are not significant, provided a substantial contribution to the model fit indices. Setting these non-significant parameters to 0, as is common in the literature, provides a solution similar to that of the Third Model noted above and does not support model fit with the data. Thus, a more advanced exercise of quasi-confirmatory approach was used, keeping these parameters included for the sake of establishing at least adequate model fit. This approach is divergent from the true, pure confirmatory approach and is considered quasi-confirmatory because it strays from steadfast rules in search of obtaining model fit. The
shift into a more fitting model with the inclusion of non-significant parameters is evidence of the relationship between statistical significance, power, and sample size (Schumacker & Lomax, 2010). It is important to note that the results obtained were a function of the quasi-confirmatory approach and not necessarily theoretically sound.

The final model was able to achieve adequate fit statistics based on the criteria set by Schumacker and Lomax (2010). Unfortunately, the Chi-square value ($\chi^2 = 90.212, df = 30, p < .001$) was significant, suggesting poor fit between the model and the data. The RMSEA (.077) was slightly above that of a well-fitting model, where criteria outline fit between the model and the data at values less than .05, but at least .08 (MacCallum, Browne, & Sugawara, 1996). Examining the remaining indices, the values for SRMR (.033), the GFI (.956), and the AGFI (.904) are much more representative of better model fit. The fit indices of the final model are presented alongside the previous models in Table 17. Overall, the inclusion of certain error covariance terms led to the enhanced fit between the data and the model.
Table 17.

**Final Confirmatory Factor Analysis Results: Fit Indices for Four Models**

<table>
<thead>
<tr>
<th>Index</th>
<th>Preliminary Model</th>
<th>Modified Model</th>
<th>Third Model</th>
<th>Final Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-Square</td>
<td>(p &lt; .001)</td>
<td>(p &lt; .001)</td>
<td>(p &lt; .001)</td>
<td>(p &lt; .001)</td>
</tr>
<tr>
<td>Root Mean Square Error of Approximation (RMSEA)</td>
<td>.207</td>
<td>.135</td>
<td>.126</td>
<td>.077</td>
</tr>
<tr>
<td>Standardized Root Mean Residual (SRMR)</td>
<td>.071</td>
<td>.054</td>
<td>.042*</td>
<td>.033*</td>
</tr>
<tr>
<td>Goodness of Fit Index (GFI)</td>
<td>.761</td>
<td>.853</td>
<td>.899</td>
<td>.956*</td>
</tr>
<tr>
<td>Adjusted Goodness of Fit Index (AGFI)</td>
<td>.684</td>
<td>.773</td>
<td>.824</td>
<td>.904*</td>
</tr>
</tbody>
</table>

*Note.* The preliminary model consists of 18 items, the modified model consists of 13 items, and the third and final models consist of 11 items. All models consist of 4 components. * indicates within the Acceptable Range based on Schumacker & Lomax (2010).

The final model consisted of four factors, all of which had moderate to strong relationships between them (Table 18). The strongest relationship between factors was found between factors 2 (i.e., addressing questions or concerns) and 3 (i.e., empathy), while the smallest factor correlation was found between factors 1 (i.e., acknowledgement and introduction) and 4 (i.e., professionalism). With the correlations ranging from .520 to .862, there is lacking of strong evidence to support discriminant validity among the factors. Table 18 below presents the factor correlation matrix.
Table 18.

Correlation Matrix for Four Factors

<table>
<thead>
<tr>
<th>Factor:</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: Acknowledgement and Introduction</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2: Addressing Questions or Concerns</td>
<td>.688</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3: Empathy</td>
<td>.774</td>
<td>.862</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>4: Professionalism</td>
<td>.520</td>
<td>.652</td>
<td>.832</td>
<td>1.000</td>
</tr>
</tbody>
</table>

*Note.* All correlations are significant (*p* < .01).

The strong positive correlation between empathy and addressing questions or concerns (*r* = .862), as well as a similarly-high value for the correlation between empathy and professionalism (*r* = .832), point to a lack of discriminant validity and lead to two potential conclusions. First, the high correlations may provide evidence of the likelihood of a single underlying factor. This single, higher-order factor may be focused on empathy-related behaviors as the defining theme of service quality. In the event these can be validated in future research using a different sample and new data to stand alone as independent factors, it may be worth consideration that people who display behaviors of empathy are more likely to engage patients or visitors in acknowledgement and introduction, addressing questions or concerns, and show professionalism during their interactions. There, empathy flows from successful engagement of patients and professionalism in the healthcare setting. Using that rationale, people who empathetic tend to extend beyond themselves and provide high quality service to patients, based on these other factors, while those who are lacking empathy provide service quality to a
lesser extent. There is minimal evidence in the existing literature to provide definitive support for either conclusion, providing an opportunity for future consideration.

Taken together, an examination of the factor loadings and error covariance terms provides evidence that the indicators, or responses to the items themselves, may not be unique measures representing the latent variable to which they ascribe. While the model has been able to demonstrate fit based on acceptable criteria of fit indices using a quasi-confirmatory approach, there is substantial inter-relatedness within the items and within the factors. Correlations among error terms, most of which are not significant, that predicate indices suggesting adequate model fit fail to provide a strong foundation for the validity and reliability of the latent variables identified. Though adequate model fit was achieved, it was at cost of several items and done based on only a small sample of respondents.

Through the principal components, exploratory, and confirmatory analyses, there were concerns with the data that needed to be managed to provide results that could be easily interpreted. The PCA was successful in identifying underlying components within the data that were aligned with the behaviors from which the items were developed. While the results of the PCA could not be confirmed in their entirety, the structure of these components was supported using a quasi-confirmatory technique. Though only a small number of items were included due to multicollinearity issues within the data, a model consisting of 11 items and four components was able to demonstrate model fit with the data statistically. That includes the error covariance terms that were not
significant, likely reducing enough measurement error to obtain fit. There remain concerns with the nature by which model and data fit were obtained, suggesting the data do not back the theory and do not meet the standards for fit based on pure confirmatory analysis techniques. The overarching structure of a model where behavioral indicators are evidence of greater behavioral components was supported by the quasi-confirmatory analysis, but need to be validated using a much larger sample and a more strict confirmatory technique.

A more detailed discussion of the limitations to the analyses that will be discussed in a later section; however, limitations caused by the sample size and item distributions lead to substantial barriers to the inclusion of all items throughout each of the analyses. Further, the analyses suggest there is a strong relationship among items and among the underlying components, leading to a cautious interpretation of the results. Working within the limitations of this investigation, the expected behaviors serve as representations of the behavioral indicators from which the items were developed.

Together, the CFA was utilized as a means of explaining the inter-relationships among the items. These inter-relationships suggest the existence of a single, fundamental underlying factor. Most notably, the items that were able to provide unique contributions to variance aligned with the behaviors from which they were developed. Given the relatively low number of items utilized in this section of analysis due to strong item inter-correlations, the Rasch analysis was conducted as a means of describing the items as they ascribe to the individual service behaviors or if they more represent a single,
unidimensional construct that is high-quality service. The Rasch analysis was intended to provide more sound conclusions regarding the dimensionality of the instrument, either providing support for the multiple factors obtained through the factor analytic techniques or to support a single underlying dimension within the items.

**Rasch Analysis**

The Rasch analysis was employed to evaluate the instrument’s psychometric properties. Specifically, the investigation sought to further establish item and scale statistics to draw conclusions pertaining to respondents and items. The findings are organized in the following manner. First, item difficulty is discussed alongside presentation of person and item fit with the Rasch model. The second section consists of a discussion on the response category usage. The third piece focuses on examination of the underlying factor structure or dimensionality of the instrument, while the final section assesses potential for differential functioning in the items (DIF). The examination of the underlying factor structure was conducted to draw clear conclusions on the presence of multiple, individual factors related attributing item responses based on the behaviors from which they were developed, or if the items can be summarized to represent a single construct of service quality.

In the following discussion, the Rasch results are presented in terms of “ability” as the person measure and “difficulty” as the item measure. When person ability is mentioned, it is in terms of the person measure where individuals with higher scores are reporting more consistent behaviors than those with lower scores. Persons who are more
“able” are those displaying the behaviors more consistently than those who are less consistent in their behaviors and does not represent their ability to be proficient in their roles. Similarly, item difficulty is interpreted as the difficulty to endorse the item at the highest level. Simply, the more difficult items are those where fewer people report a high frequency, and items that are “easier” indicate most respondents engage in the behavior much more consistently, providing higher ratings. Given what the survey is assessing, person ability and item difficulty must be conceptualized this way when using the Rasch language.

**Scale and item statistics.** The scale and item statistics is broken down into two components: person and item analyses. Here, the person statistics are presented first, followed by the item results. This section concludes with results combining person and item statistics.

**Person statistics.** The initial analysis utilized all 334 participants that provided responses to each of the 30 rating scale items. The maximum extreme score (150) was achieved by 78 respondents (23.4%), who were automatically excluded from the analysis. Thus, the remaining analyses are based on the results of 256 participants. Table 19 provides the summary statistics for persons.
Table 19.

Summary of 256 Measured (Non-Extreme) Persons (30 Measured Items)

<table>
<thead>
<tr>
<th>Value</th>
<th>Total Score</th>
<th>Measure</th>
<th>Model Error</th>
<th>Infit MNSQ</th>
<th>ZSTD</th>
<th>Outfit MNSQ</th>
<th>ZSTD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>140.80</td>
<td>3.03</td>
<td>0.50</td>
<td>1.11</td>
<td>0.30</td>
<td>0.93</td>
<td>0.00</td>
</tr>
<tr>
<td>SD</td>
<td>8.70</td>
<td>1.18</td>
<td>0.23</td>
<td>0.56</td>
<td>1.10</td>
<td>0.53</td>
<td>0.90</td>
</tr>
<tr>
<td>Maximum</td>
<td>149.00</td>
<td>5.05</td>
<td>1.01</td>
<td>4.12</td>
<td>4.50</td>
<td>5.09</td>
<td>4.40</td>
</tr>
<tr>
<td>Minimum</td>
<td>95.00</td>
<td>0.00</td>
<td>0.18</td>
<td>0.38</td>
<td>-2.50</td>
<td>0.29</td>
<td>-2.40</td>
</tr>
</tbody>
</table>

Real RMSE: .59  True SD: 1.02  Separation: 1.75  Person Reliability: .75
Model RMSE: .55  True SD: 1.04  Separation: 1.91  Person Reliability: .78
S.E. of Person Mean: .07

The Total Score column presents person statistics on the entire instrument, calculated by adding all of the subject’s responses. This is useful in evaluating the subjects’ average responses relative to the scale’s highest possible score, which is 150 in this case. It is apparent again that subjects seemed to consistently endorse higher values, as the mean (140.80) approaches the highest possible score on the instrument.

Person ability estimates (i.e., “Measure”) ranged from 0.00 to 5.05 on the logit scale, with a mean of 3.03. This is near the high end, suggesting person ability is greater than captured adequately by the items. Further reliability of person estimates was found to be .75, which is considered moderate given its range from 0 to 1 and interpretation similar to Coefficient Alpha (Bond & Fox, 2007). In regard to persons, separation was 1.75 for the data at hand (i.e., real), and was 1.91 when the data have no misfit to the model (i.e., model). These values were expected to be much greater than one, indicating
the persons are not spread vastly over the logit scale of ability. Lastly, the real (.59) and model (.55) Root Mean Square Error (RMSE) values are close, representing evidence of little error in the data.

Person fit summary statistics suggest there are persons who are mis-fitting with the measurement model. The range of mean-square (MNSQ) and standardized (ZSTD) fit statistics for both infit and outfit provide values outside the commonly-accepted ranges for good fit. Persons and items are typically concluded to demonstrate fit when mean-square values are between .5 and 1.5 and standardized fit statistics range from -2 to 2 (de Ayala, 2009). Based on the range provided in Table 19, infit and outfit each show values well beyond the common expectations for both mean-square and standardized fit. A further investigation into person misfit is warranted.

Upon examination of the Person Misfit Order (Table 20), the mean-square and standardized fit statistics were used to identify persons providing excess misfit with the model that could be removed. Taking into account the interpretation of mean-square fit statistics (Wright & Linacre, 1994) and other criteria for standardized fit values (e.g., de Ayala, 2009), persons were considered candidates for removal if they obtained a standardized value outside the range from -2.00 to 2.00, or if they obtained mean-square fit statistics outside the range from .5 to 1.75, on either infit or outfit. In the event a person obtained either infit or outfit values outside this range, they were identified and removed. The inclusion of persons with mean-square fit statistics from 1.5 to 1.75 may include persons not productive for measurement, but not degrading to the results (Wright
& Linacre, 1994). The criterion of an infit or outfit mean square greater than 1.75 was adopted to prevent removing too many respondents based on one mean-square value alone; those with mean-square values in the 1.5 to 1.75 range were removed if either of the standardized values exceeded 2.0. A total of 34 respondents were removed from the analysis based on extreme misfit scores. The analyses proceeded using the sample of 222 respondents.
### Table 20.

**Person Statistics: Misfit Order (30 Measured Items)**

<table>
<thead>
<tr>
<th>Person</th>
<th>Total Score</th>
<th>Measure</th>
<th>S.E.</th>
<th>Model MNSQ</th>
<th>ZSTD</th>
<th>Infit MNSQ</th>
<th>ZSTD</th>
<th>Outfit MNSQ</th>
<th>ZSTD</th>
</tr>
</thead>
<tbody>
<tr>
<td>103</td>
<td>144</td>
<td>3.10</td>
<td>0.44</td>
<td>3.31</td>
<td>3.80</td>
<td>5.09</td>
<td>4.40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>66</td>
<td>146</td>
<td>3.56</td>
<td>0.53</td>
<td>4.12</td>
<td>3.90</td>
<td>2.17</td>
<td>1.60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>164</td>
<td>146</td>
<td>3.56</td>
<td>0.53</td>
<td>3.98</td>
<td>3.80</td>
<td>1.79</td>
<td>1.20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>323</td>
<td>147</td>
<td>3.88</td>
<td>0.60</td>
<td>1.78</td>
<td>1.30</td>
<td>3.32</td>
<td>2.20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>167</td>
<td>125</td>
<td>1.22</td>
<td>0.24</td>
<td>3.16</td>
<td>4.50</td>
<td>2.67</td>
<td>3.90</td>
<td></td>
<td></td>
</tr>
<tr>
<td>177</td>
<td>135</td>
<td>1.93</td>
<td>0.30</td>
<td>3.15</td>
<td>4.10</td>
<td>2.17</td>
<td>2.70</td>
<td></td>
<td></td>
</tr>
<tr>
<td>145</td>
<td>132</td>
<td>1.68</td>
<td>0.28</td>
<td>3.02</td>
<td>4.00</td>
<td>2.96</td>
<td>4.20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>294</td>
<td>144</td>
<td>3.10</td>
<td>0.44</td>
<td>3.02</td>
<td>3.40</td>
<td>1.36</td>
<td>0.80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>143</td>
<td>2.91</td>
<td>0.42</td>
<td>2.84</td>
<td>3.40</td>
<td>1.56</td>
<td>1.20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>245</td>
<td>139</td>
<td>2.34</td>
<td>0.34</td>
<td>2.77</td>
<td>3.50</td>
<td>1.59</td>
<td>1.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>149</td>
<td>5.05</td>
<td>1.01</td>
<td>1.06</td>
<td>0.40</td>
<td>2.71</td>
<td>1.30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>189</td>
<td>148</td>
<td>4.32</td>
<td>0.73</td>
<td>2.12</td>
<td>1.50</td>
<td>2.62</td>
<td>1.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>182</td>
<td>137</td>
<td>2.12</td>
<td>0.32</td>
<td>2.50</td>
<td>3.20</td>
<td>1.60</td>
<td>1.60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>274</td>
<td>131</td>
<td>1.61</td>
<td>0.27</td>
<td>2.48</td>
<td>3.20</td>
<td>1.84</td>
<td>2.20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>146</td>
<td>3.56</td>
<td>0.53</td>
<td>2.40</td>
<td>2.30</td>
<td>1.27</td>
<td>0.60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>297</td>
<td>146</td>
<td>3.56</td>
<td>0.53</td>
<td>2.28</td>
<td>2.10</td>
<td>1.00</td>
<td>0.20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>131</td>
<td>1.61</td>
<td>0.27</td>
<td>2.24</td>
<td>2.80</td>
<td>1.83</td>
<td>2.20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>133</td>
<td>135</td>
<td>1.93</td>
<td>0.30</td>
<td>2.11</td>
<td>2.50</td>
<td>1.51</td>
<td>1.40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>265</td>
<td>148</td>
<td>4.32</td>
<td>0.73</td>
<td>2.06</td>
<td>1.40</td>
<td>1.63</td>
<td>0.90</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>142</td>
<td>2.75</td>
<td>0.39</td>
<td>1.99</td>
<td>2.20</td>
<td>0.99</td>
<td>0.10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>168</td>
<td>146</td>
<td>3.56</td>
<td>0.53</td>
<td>1.99</td>
<td>1.80</td>
<td>1.34</td>
<td>0.70</td>
<td></td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>130</td>
<td>1.53</td>
<td>0.26</td>
<td>1.69</td>
<td>1.80</td>
<td>1.98</td>
<td>2.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>121</td>
<td>141</td>
<td>2.60</td>
<td>0.37</td>
<td>1.96</td>
<td>2.20</td>
<td>1.28</td>
<td>0.80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>282</td>
<td>148</td>
<td>4.32</td>
<td>0.73</td>
<td>1.05</td>
<td>0.30</td>
<td>1.96</td>
<td>1.10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>61</td>
<td>129</td>
<td>1.46</td>
<td>0.26</td>
<td>1.94</td>
<td>2.30</td>
<td>1.74</td>
<td>2.10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>261</td>
<td>95</td>
<td>0.00</td>
<td>0.18</td>
<td>1.67</td>
<td>2.40</td>
<td>1.92</td>
<td>2.80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>263</td>
<td>143</td>
<td>2.91</td>
<td>0.42</td>
<td>1.33</td>
<td>0.90</td>
<td>1.87</td>
<td>1.60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>95</td>
<td>122</td>
<td>1.06</td>
<td>0.23</td>
<td>0.45</td>
<td>-2.20</td>
<td>0.54</td>
<td>-1.80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>150</td>
<td>108</td>
<td>0.45</td>
<td>0.19</td>
<td>0.54</td>
<td>-2.00</td>
<td>0.46</td>
<td>-2.40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>278</td>
<td>125</td>
<td>1.22</td>
<td>0.24</td>
<td>0.40</td>
<td>-2.30</td>
<td>0.51</td>
<td>-1.90</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 20. (continued)

<table>
<thead>
<tr>
<th>Person</th>
<th>Total Score</th>
<th>Measure</th>
<th>Model S.E.</th>
<th>Infit MNSQ</th>
<th>Outfit ZSTD</th>
<th>Person Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>152</td>
<td>123</td>
<td>1.11</td>
<td>0.23</td>
<td>0.47</td>
<td>-2.10</td>
<td>0.49</td>
</tr>
<tr>
<td>201</td>
<td>120</td>
<td>0.96</td>
<td>0.22</td>
<td>0.46</td>
<td>-2.20</td>
<td>0.48</td>
</tr>
<tr>
<td>91</td>
<td>131</td>
<td>1.61</td>
<td>0.27</td>
<td>0.38</td>
<td>-2.30</td>
<td>0.45</td>
</tr>
<tr>
<td>216</td>
<td>119</td>
<td>0.91</td>
<td>0.22</td>
<td>0.41</td>
<td>-2.50</td>
<td>0.44</td>
</tr>
<tr>
<td>Mean</td>
<td>142.9</td>
<td>3.79</td>
<td>0.81</td>
<td>1.11</td>
<td>0.30</td>
<td>0.93</td>
</tr>
<tr>
<td>SD</td>
<td>8.6</td>
<td>1.72</td>
<td>0.60</td>
<td>0.56</td>
<td>1.10</td>
<td>0.53</td>
</tr>
</tbody>
</table>

Note. Not all person statistics are shown, only the most mis-fitting are provided.

The removal of the extreme cases was met by slight increases in person separation (1.75 to 1.78) and person reliability (.75 to .76). The mean person ability estimate was more than half a logit higher at 3.66, compared with 3.03 prior to person removal. The updated person summary statistics are provided in Table 21 below.

Table 21.

Summary of 222 Measured (Non-Extreme) Persons (30 Measured Items)

<table>
<thead>
<tr>
<th>Value</th>
<th>Total Score</th>
<th>Measure</th>
<th>Model Error</th>
<th>Infit MNSQ</th>
<th>Outfit MNSQ</th>
<th>ZSTD</th>
<th>ZSTD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>141.70</td>
<td>3.66</td>
<td>0.52</td>
<td>1.03</td>
<td>0.20</td>
<td>0.86</td>
<td>0.00</td>
</tr>
<tr>
<td>SD</td>
<td>7.60</td>
<td>1.21</td>
<td>0.23</td>
<td>0.29</td>
<td>0.80</td>
<td>0.35</td>
<td>0.70</td>
</tr>
<tr>
<td>Maximum</td>
<td>149.00</td>
<td>5.64</td>
<td>1.02</td>
<td>1.86</td>
<td>2.40</td>
<td>2.05</td>
<td>2.70</td>
</tr>
<tr>
<td>Minimum</td>
<td>96.00</td>
<td>0.05</td>
<td>0.20</td>
<td>0.51</td>
<td>-1.90</td>
<td>0.28</td>
<td>-1.60</td>
</tr>
</tbody>
</table>

Real RMSE: .59  True SD: 1.05  Separation: 1.78  Person Reliability: .76
Model RMSE: .55  True SD: 1.06  Separation: 1.86  Person Reliability: .78
S.E. of Person Mean: .08
Person Raw Score-To-Measure Correlation = .92
Cronbach Alpha (KR-20) Person Raw Score "Test" Reliability = .89
The obtained raw score-to-measure correlation (.92) is quite strong, as it represents the correlation between person raw scores and person ability estimates on the corresponding measure. This represents the extent to which people who obtain higher raw scores are rated as more able than respondents with lower total scores. This value should be as close as possible to 1, suggesting the people with the highest ability obtain the highest scores (Linacre, 2012b). Similarly, the Alpha for these items was found to be quite high as well (.89), supplementing the data provided in the item analysis. Based on the infit and outfit ranges provided, there seem to still be some outlying respondents. Prior to removal of any additional persons, the items will be investigated to evaluate their fit with the model.

**Item statistics.** The items were centered so the mean was 0 with an obtained standard deviation of .90. The item difficulties spanned from -1.76 to 1.65, suggesting the 30 items were clustered near the mean in terms of relative difficulty. Item separation was greater than for persons, with items achieving a real separation of 4.83 and a model separation of 4.98. There was less error involved in the items as the RMSE values were .18 for both real and model values, and reliability was much higher (.96) with real and model values being equal. The item raw score-to-measure correlation (-.97) approaches -1, indicating less difficult items were met with correct responses. Overall, the Chi-square value suggests the data is not misfit with the Rasch model ($\chi^2 = 6590.56$, $df = 6406$, $p > .05$). Table 22 provides the item summary statistics.
Table 22.

Summary of 30 Measured Items (222 Measured Persons)

<table>
<thead>
<tr>
<th>Value</th>
<th>Total Score</th>
<th>Measure</th>
<th>Model Error</th>
<th>Infit MNSQ</th>
<th>ZSTD</th>
<th>Outfit MNSQ</th>
<th>ZSTD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>1438.30</td>
<td>0.00</td>
<td>0.17</td>
<td>0.99</td>
<td>0.00</td>
<td>0.86</td>
<td>-0.30</td>
</tr>
<tr>
<td>SD</td>
<td>38.50</td>
<td>0.90</td>
<td>0.05</td>
<td>0.24</td>
<td>1.60</td>
<td>0.35</td>
<td>1.70</td>
</tr>
<tr>
<td>Maximum</td>
<td>1488.00</td>
<td>1.65</td>
<td>0.30</td>
<td>1.60</td>
<td>4.40</td>
<td>1.55</td>
<td>3.80</td>
</tr>
<tr>
<td>Minimum</td>
<td>1345.00</td>
<td>-1.76</td>
<td>0.11</td>
<td>0.62</td>
<td>-3.00</td>
<td>0.30</td>
<td>-2.80</td>
</tr>
</tbody>
</table>

Real RMSE: .18  True SD: .88  Separation: 4.83  Person Reliability: .96
Model RMSE: .18 True SD: .88  Separation: 4.98  Person Reliability: .96
S.E. of Item Mean: .17
Item Raw Score-To-Measure Correlation = -.97
Log-Likelihood Chi-Square = 6590.56 (p > .05)

The range of infit and outfit values provides evidence of some items not fitting adequately to the Rasch model. The same criteria apply for infit and outfit of items as they do persons, so a more in-depth look into mis-fitting items was needed. Using the Item Misfit Order table (Table 23), 11 items (i.e., Items 1, 3, 12, 13, 18, 17, 20, 21, 22, 25, and 30) were identified as misfit with the data based on the standardized fit statistics (i.e., infit or outfit values in excess of 2.00 or -2.00). Outfit statistics are un-weighted and are more sensitive to outlying scores; leading to increased importance placed on the infit values (Bond & Fox, 2007). Therefore, items that are problematic to the measurement model were identified using infit mean-square and standardized fit statistics, including outfit values as supplemental.
Table 23.

Item Statistics: Misfit Order (222 Measured Persons)

<table>
<thead>
<tr>
<th>Item</th>
<th>Total Score</th>
<th>Measure</th>
<th>Model S.E.</th>
<th>Infit MNSQ</th>
<th>ZSTD</th>
<th>Outfit MNSQ</th>
<th>ZSTD</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>1351</td>
<td>1.59</td>
<td>0.11</td>
<td>1.60</td>
<td>4.4</td>
<td>1.27</td>
<td>2.1</td>
</tr>
<tr>
<td>1</td>
<td>1370</td>
<td>1.36</td>
<td>0.11</td>
<td>1.22</td>
<td>1.8</td>
<td>1.55</td>
<td>3.8</td>
</tr>
<tr>
<td>30</td>
<td>1412</td>
<td>0.74</td>
<td>0.13</td>
<td>1.40</td>
<td>3.0</td>
<td>1.54</td>
<td>3.0</td>
</tr>
<tr>
<td>20</td>
<td>1428</td>
<td>0.45</td>
<td>0.14</td>
<td>1.39</td>
<td>2.8</td>
<td>1.50</td>
<td>2.5</td>
</tr>
<tr>
<td>3</td>
<td>1435</td>
<td>0.30</td>
<td>0.15</td>
<td>1.42</td>
<td>2.9</td>
<td>1.17</td>
<td>0.9</td>
</tr>
<tr>
<td>18</td>
<td>1460</td>
<td>-0.35</td>
<td>0.18</td>
<td>0.80</td>
<td>-1.3</td>
<td>0.49</td>
<td>-2.3</td>
</tr>
<tr>
<td>13</td>
<td>1472</td>
<td>-0.79</td>
<td>0.21</td>
<td>0.76</td>
<td>-1.4</td>
<td>0.44</td>
<td>-2.1</td>
</tr>
<tr>
<td>21</td>
<td>1482</td>
<td>-1.31</td>
<td>0.25</td>
<td>0.76</td>
<td>-1.1</td>
<td>0.30</td>
<td>-2.2</td>
</tr>
<tr>
<td>22</td>
<td>1471</td>
<td>-0.75</td>
<td>0.20</td>
<td>0.74</td>
<td>-1.5</td>
<td>0.42</td>
<td>-2.2</td>
</tr>
<tr>
<td>25</td>
<td>1479</td>
<td>-1.13</td>
<td>0.23</td>
<td>0.73</td>
<td>-1.4</td>
<td>0.32</td>
<td>-2.3</td>
</tr>
<tr>
<td>12</td>
<td>1449</td>
<td>-0.03</td>
<td>0.16</td>
<td>0.62</td>
<td>-3.0</td>
<td>0.48</td>
<td>-2.8</td>
</tr>
<tr>
<td>Mean</td>
<td>1438.3</td>
<td>0.00</td>
<td>0.17</td>
<td>0.99</td>
<td>0.0</td>
<td>0.86</td>
<td>-0.3</td>
</tr>
<tr>
<td>SD</td>
<td>38.5</td>
<td>0.90</td>
<td>0.05</td>
<td>0.24</td>
<td>1.6</td>
<td>0.35</td>
<td>1.7</td>
</tr>
</tbody>
</table>

There are five items determined to be problematic based on the standardized infit values (i.e., Items 17, 30, 20, 3, and 12). Item 3 was the only one of the five identified to not exceed the cutoff of 2.00 on the outfit value as well. All infit mean-square values are within the range of acceptance established for persons (i.e., between .5 and 1.75). The point-measure correlations range from .30 to .62, which indicates a moderate relationship between the item and measure values and no evidence of substantial misfit. Overall, most of the items were able to demonstrate adequate fit with the Rasch model and all were used throughout further analysis. Items were not immediately removed as the research investigated the dimensionality and potential for differential functioning of the items, which could contribute to problematic fit statistics.
**Person ability and item difficulty.** As previously noted, person measure estimates ranged from .05 to 5.64 logits, with a mean of 3.66. Item means ranged from -1.76 to 1.65 logits with the mean centered on 0. A visual depiction of the item and person distributions is provided via the person-item map (Figure 3). When placed on the same scale, it becomes apparent that respondents consistently out-performed the items. There is minimal overlap between the ranges covered by persons and items independently, presenting concerns relevant to the developed items. Collectively, respondents found the items to be easy to endorse, providing responses to a large number of items with very high values.
Figure 3. The map of persons and items for the service behavior instrument. Items are distributed on the right side of the vertical line, identified by their question number, and persons are distributed on the left in aggregate. Each “#” sign represents 6 respondents while each “.” represents 1 to 5 respondents.
Concerns exist relative to the large differences between person ability and item difficulty. Essentially, the items do not adequately capture the differences in expected service behaviors among the sample. The range of item difficulties well below most person ability estimates reflects the items’ inability to distinguish among higher-performing individuals. Quite simply, there was no item that served to discriminate very well among those respondents who reported frequent exhibitions of the target behaviors. The lack of truly challenging items led to a large group of respondents outside the scope of the difficulties present in the items.

**Category structure and usage.** The next piece of the Rasch analysis examined the usage of response categories via threshold values. Much like estimates were made on the relative difficulty of items, estimates were made on the difficulties predicted for choosing one response category over another. These values are called thresholds or step calibrations, and represent the transition point at which the next response category becomes more likely (de Ayala, 2009). The threshold values are expected to increase across the ordered response categories and should be spread along the logit scale, with 1.4 to 5 logits between obtained threshold values (Bond & Fox, 2007). Category structure investigation was done to provide insight into response option usage and to draw conclusions relevant to the responses provided on the instrument.

The frequencies of category usage and threshold values are provided in Table 24. As previously established, responses with values of “1” (i.e., “Never or Almost Never”) and “2” (“Rarely”) were utilized 10 and 17 times, respectively, comprising less than 1% of all responses. Participants endorsed “5” (i.e., “Always or Almost Always”) 77% of the
time. Since there are five possible responses, four threshold values were calculated. Due to the under-utilization of a “2” response, the thresholds are out of order and a response of “3” was more likely, leading to no point along the range of theta that “2” was the most likely response. A visual depiction of the category probabilities is provided in Figure 4.

Table 24.

<table>
<thead>
<tr>
<th>Category Label</th>
<th>Observed Count</th>
<th>%</th>
<th>Observed Average</th>
<th>Sample Expect</th>
<th>Infit MNSQ</th>
<th>Outfit MNSQ</th>
<th>Andrich Threshold</th>
<th>Category Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>0</td>
<td>0.64</td>
<td>-0.19</td>
<td>1.62</td>
<td>1.92</td>
<td>None</td>
<td>(-2.32)</td>
</tr>
<tr>
<td>2</td>
<td>17</td>
<td>0</td>
<td>1.07</td>
<td>0.53</td>
<td>1.38</td>
<td>1.52</td>
<td>-0.36</td>
<td>-1.23</td>
</tr>
<tr>
<td>3</td>
<td>221</td>
<td>3</td>
<td>1.41</td>
<td>1.35</td>
<td>1.08</td>
<td>1.05</td>
<td>-1.64</td>
<td>-0.33</td>
</tr>
<tr>
<td>4</td>
<td>1317</td>
<td>20</td>
<td>2.42</td>
<td>2.49</td>
<td>0.98</td>
<td>0.75</td>
<td>0.1</td>
<td>1.08</td>
</tr>
<tr>
<td>5</td>
<td>5095</td>
<td>77</td>
<td>4.09</td>
<td>4.07</td>
<td>1.02</td>
<td>1.01</td>
<td>1.9</td>
<td>-3.11</td>
</tr>
</tbody>
</table>
Figure 4. Category probability curves. The category probability curves indicate the probability of a response for all items. The curves depict the likelihood of response category selection (Y-axis) by the person-item measure (X-axis). The category use above shows response category 2, “Rarely”, under-represented and participants likelihood to endorse a higher frequency of the behaviors assessed.

Overall, the category probability curves and threshold values support the evidence provided above; namely, participants were much more likely to endorse a higher frequency of the behaviors. With participants endorsing the target behaviors at such a high frequency, the items are identified to be less difficult and participants are rated as having higher ability. With about 77% of respondents endorsing the highest frequency across items and more than 97% of the responses being accounted for in the top two
categories, the response structure is difficult to establish when examining the lower end of the response scale.

To rectify the disordered thresholds and under-used response categories near the lower end of the response scale, response options 1 and 2 were collapsed into one category (i.e., “Never or almost never” and “Rarely” combine to “Never or rarely”). This pools the frequencies of the under-used responses with the intention of a more prominent item characteristic curve and more appropriate threshold values (Table 25). Upon combining the categories, the 27 responses still constituted less than 1% of the observed responses. However, the new threshold values (Table 25) more consistently approximate the criteria of 1.4 logits, at minimum, between thresholds (Bond & Fox, 2007). Further, the thresholds are ordered properly and lower response options are more likely for items deemed easier. Table 25 provides the updated category structure values once categories were collapsed. The updated category probability curves are provided in Figure 5 below.

Table 25.

<table>
<thead>
<tr>
<th>Category Label</th>
<th>Observed Count</th>
<th>Observed %</th>
<th>Sample Average</th>
<th>Sample Expect</th>
<th>Infit MNSQ</th>
<th>Outfit MNSQ</th>
<th>Andrich Threshold</th>
<th>Category Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>27</td>
<td>0</td>
<td>.75</td>
<td>0.01</td>
<td>1.45</td>
<td>1.49</td>
<td>None</td>
<td>-2.80</td>
</tr>
<tr>
<td>2</td>
<td>221</td>
<td>3</td>
<td>1.14</td>
<td>1.07</td>
<td>1.08</td>
<td>1.06</td>
<td>-1.55</td>
<td>-0.91</td>
</tr>
<tr>
<td>3</td>
<td>1317</td>
<td>20</td>
<td>2.20</td>
<td>2.27</td>
<td>0.98</td>
<td>0.75</td>
<td>-0.14</td>
<td>0.86</td>
</tr>
<tr>
<td>4</td>
<td>5095</td>
<td>77</td>
<td>3.89</td>
<td>3.87</td>
<td>1.03</td>
<td>1.01</td>
<td>1.69</td>
<td>2.90</td>
</tr>
</tbody>
</table>

*Note.* Category Label 1 refers to “Never or Rarely”, Label 2 refers to “Sometimes”, Label 3 refers to “Frequently”, and Label 4 refers to “Always or almost always”.
Figure 5. Category probability curves updated with the collapsed categories. The category probability curves indicate the probability of a response for all items. The curves depict the likelihood of response category selection (Y-axis) by the person-item measure (X-axis). The category use above shows good representation and ordered thresholds of response categories.

**Underlying factor structure and dimensionality.** The next analysis examined the underlying factor structure or dimensionality of the instrument. One assumption made in IRT analyses is that the instrument is unidimensional, or that the items represent only one latent trait at a time (Linacre, 1998). This assumption was tested using PCA of the Rasch residual values. This analysis provides a breakdown of variance explained by
different components in the data (i.e., variance explained by persons, items, and unexplained variance), which is presented in Eigenvalue units and percentages.

The amount of variance explained by the measures was found to be 32.8%, which is considered limited as it is subject to the spread of persons and items (Linacre, 2012a). In this case, limited variance explained is not atypical as the population has a narrow range of ability (i.e., they have all been through the training). The obtained 32.8% of variance explained is similar to the modeled 36.5%, so the data as a whole are slightly under-fit to the Rasch model. The unexplained variance, 67.2%, is slightly more than expected if the data fit the Rasch model perfectly (63.5%). This unexplained variance was broken down using a PCA of the Rasch residuals in subsequent contrasts to evaluate the strength of underlying dimensions. A scree plot of the variance components and contrasts is provided in Figure 6, while Table 26 presents the component variance Eigenvalues and percentages for both the obtained data (i.e., Observed) and modeled (i.e., Expected).
**Figure 6.** Scree plot of standardized residual variance. This log-scaled scree plot shows the respective percentages of the variance components. They are ordered as follows: Total Variance, Measure Variance, Person Variance, Item Variance, and Unexplained Variance, followed by each of the five contrasts (i.e., U1-U5).
Table 26.

*Standardized Residual Variance (in Eigenvalue units)*

<table>
<thead>
<tr>
<th></th>
<th>Observed</th>
<th></th>
<th>Expected</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Eigenvalue</td>
<td>Percent</td>
<td>Percent</td>
</tr>
<tr>
<td>Total Raw Variance in Observations</td>
<td>44.6</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>Raw Variance Explained by Measures</td>
<td>14.6</td>
<td>32.8</td>
<td>36.5</td>
</tr>
<tr>
<td>Raw Variance Explained by Persons</td>
<td>8.5</td>
<td>19.0</td>
<td>21.2</td>
</tr>
<tr>
<td>Raw Variance Explained by Items</td>
<td>6.1</td>
<td>13.8</td>
<td>15.3</td>
</tr>
<tr>
<td>Raw Unexplained Variance (total)</td>
<td>30.0</td>
<td>67.2</td>
<td>63.5</td>
</tr>
<tr>
<td>Unexplained Variance in 1st Contrast</td>
<td>2.4</td>
<td>5.5</td>
<td></td>
</tr>
<tr>
<td>Unexplained Variance in 2nd Contrast</td>
<td>2.1</td>
<td>4.7</td>
<td></td>
</tr>
<tr>
<td>Unexplained Variance in 3rd Contrast</td>
<td>1.9</td>
<td>4.2</td>
<td></td>
</tr>
<tr>
<td>Unexplained Variance in 4th Contrast</td>
<td>1.7</td>
<td>3.8</td>
<td></td>
</tr>
<tr>
<td>Unexplained Variance in 5th Contrast</td>
<td>1.5</td>
<td>3.4</td>
<td></td>
</tr>
</tbody>
</table>

While there remained a substantial amount of unexplained variance (67.2%) in the measure, the first contrast provided only 5.5% of the total variance in the observations, or 8.1% of the unexplained variance. This was the largest contrast possible, yielding an Eigenvalue of 2.4. This can be equated to the strength of 2.4 items in the contrast (Linacre, 2012a). Examining the loading values on the first dimension of residuals (Table 27), only five items achieved loadings greater than .40, three of them positive (i.e., Items 21, 25, and 29) and two of them negative (i.e., Items 1 and 17). The loading values represent the loadings that are off the Rasch dimension, with those greater than .40 are substantial (Bond & Fox, 2007), while items that cluster together on the Component Factor Plot (Figure 7) are those that may be further investigated for disturbances in measurement (Wright & Stone, 2004). The next step in determining dimensionality
involves a closer look into these items to establish if they are sub-dimensions or if they constitute random effects in the data (Linacre, 2012a).

Table 2.

*Standardized Residual Loadings for Items in Contrast 1 (Sorted by Loading)*

<table>
<thead>
<tr>
<th>Loading</th>
<th>Measure</th>
<th>Infit MNSQ</th>
<th>Outfit MNSQ</th>
<th>Entry</th>
<th>Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.61</td>
<td>-1.31</td>
<td>0.76</td>
<td>0.30</td>
<td>A</td>
<td>Item 21</td>
</tr>
<tr>
<td>0.52</td>
<td>-1.13</td>
<td>0.73</td>
<td>0.32</td>
<td>B</td>
<td>Item 25</td>
</tr>
<tr>
<td>0.50</td>
<td>-0.83</td>
<td>0.81</td>
<td>0.57</td>
<td>C</td>
<td>Item 29</td>
</tr>
<tr>
<td>0.38</td>
<td>0.09</td>
<td>0.90</td>
<td>0.74</td>
<td>D</td>
<td>Item 10</td>
</tr>
<tr>
<td>0.37</td>
<td>-0.93</td>
<td>0.86</td>
<td>0.75</td>
<td>E</td>
<td>Item 28</td>
</tr>
<tr>
<td>0.34</td>
<td>-1.02</td>
<td>0.82</td>
<td>0.47</td>
<td>F</td>
<td>Item 24</td>
</tr>
<tr>
<td>0.32</td>
<td>-0.79</td>
<td>0.76</td>
<td>0.44</td>
<td>G</td>
<td>Item 13</td>
</tr>
<tr>
<td>0.32</td>
<td>0.02</td>
<td>0.87</td>
<td>0.63</td>
<td>H</td>
<td>Item 27</td>
</tr>
<tr>
<td>0.29</td>
<td>-0.03</td>
<td>0.62</td>
<td>0.48</td>
<td>I</td>
<td>Item 12</td>
</tr>
<tr>
<td>0.25</td>
<td>-0.67</td>
<td>1.34</td>
<td>1.07</td>
<td>J</td>
<td>Item 19</td>
</tr>
<tr>
<td>0.25</td>
<td>-0.75</td>
<td>0.74</td>
<td>0.42</td>
<td>K</td>
<td>Item 22</td>
</tr>
<tr>
<td>0.14</td>
<td>-0.67</td>
<td>1.00</td>
<td>0.93</td>
<td>L</td>
<td>Item 16</td>
</tr>
<tr>
<td>0.07</td>
<td>0.64</td>
<td>0.96</td>
<td>1.11</td>
<td>M</td>
<td>Item 15</td>
</tr>
<tr>
<td>0.07</td>
<td>0.74</td>
<td>1.40</td>
<td>1.54</td>
<td>N</td>
<td>Item 30</td>
</tr>
<tr>
<td>0.05</td>
<td>0.94</td>
<td>0.91</td>
<td>0.89</td>
<td>O</td>
<td>Item 9</td>
</tr>
<tr>
<td>0.03</td>
<td>0.32</td>
<td>0.98</td>
<td>0.94</td>
<td>o</td>
<td>Item 7</td>
</tr>
<tr>
<td>0.01</td>
<td>-0.23</td>
<td>1.14</td>
<td>1.08</td>
<td>n</td>
<td>Item 2</td>
</tr>
<tr>
<td>0.01</td>
<td>-1.19</td>
<td>0.88</td>
<td>0.75</td>
<td>m</td>
<td>Item 5</td>
</tr>
<tr>
<td>0.01</td>
<td>0.62</td>
<td>0.86</td>
<td>0.87</td>
<td>l</td>
<td>Item 8</td>
</tr>
<tr>
<td>0.01</td>
<td>0.49</td>
<td>1.06</td>
<td>0.90</td>
<td>k</td>
<td>Item 23</td>
</tr>
<tr>
<td>-0.44</td>
<td>1.36</td>
<td>1.22</td>
<td>1.55</td>
<td>a</td>
<td>Item 1</td>
</tr>
<tr>
<td>-0.43</td>
<td>1.59</td>
<td>1.60</td>
<td>1.27</td>
<td>b</td>
<td>Item 17</td>
</tr>
<tr>
<td>-0.35</td>
<td>1.01</td>
<td>1.10</td>
<td>1.11</td>
<td>c</td>
<td>Item 14</td>
</tr>
<tr>
<td>-0.30</td>
<td>0.30</td>
<td>1.42</td>
<td>1.17</td>
<td>d</td>
<td>Item 3</td>
</tr>
<tr>
<td>-0.29</td>
<td>0.84</td>
<td>1.06</td>
<td>1.18</td>
<td>e</td>
<td>Item 4</td>
</tr>
<tr>
<td>-0.10</td>
<td>-1.76</td>
<td>0.83</td>
<td>0.46</td>
<td>f</td>
<td>Item 6</td>
</tr>
<tr>
<td>-0.07</td>
<td>-0.35</td>
<td>0.80</td>
<td>0.49</td>
<td>g</td>
<td>Item 18</td>
</tr>
<tr>
<td>-0.06</td>
<td>0.60</td>
<td>0.84</td>
<td>0.74</td>
<td>h</td>
<td>Item 11</td>
</tr>
<tr>
<td>-0.05</td>
<td>1.65</td>
<td>1.03</td>
<td>1.14</td>
<td>i</td>
<td>Item 26</td>
</tr>
<tr>
<td>-0.02</td>
<td>0.45</td>
<td>1.39</td>
<td>1.50</td>
<td>j</td>
<td>Item 20</td>
</tr>
</tbody>
</table>

*Note.* “Entry” column refers to the item’s identification in Figure 6.
Figure 7. Component factor plot. This plot displays items, sorted by loading along the y-axis and difficulty along the x-axis, based on the first contrast. The items appear to be remotely clustered by measurement levels or difficulty.

The next logical step is to establish if the clusters of items by loading values resemble unique behaviors. The three items with factor loadings exceeding .40 are Item 21 (i.e., “I greet others warmly.”), Item 25 (i.e., “I do my best to show compassion for patients or families.”), and Item 29 (i.e., “I recognize that each patient/visitor has a unique experience and circumstances.”). The first reflects welcoming while the second is indicative of empathy and the third could be considered global indicator of service quality. The two items with the largest negative loadings were Item 1 (i.e., “I am knowledgeable about patient care.”) and Item 17 (i.e., “I am forthcoming in sharing what
patients can expect from any procedure.”), which could share some level of “professionalism” with one another. Most importantly, there is nothing easily apparent to distinguish the latter two from the former three; there seems to be no logical reason to split the items into separate sub-dimensions for analysis. Further, four of the five items were identified as problematic based on fit statistics and were included in this analysis to determine if misfit was due to another underlying dimension. Very high standardized values of infit and outfit were found for Items 1 and 17, while Items 21 and 25 were found to have very strong, negative standardized outfit statistics. The results of the dimensionality investigation suggest there may be a secondary dimension present within the data. However, the items providing evidence of a secondary, underlying dimension are those not fit with the Rasch model based on their fit statistics. Further, there is no relationship between the items based on content, so a secondary dimension is not likely.

**Differential item functioning.** The items were also subject to an examination of differential item functioning (DIF). The analyses follow two grouping variables: gender and the participant’s involvement with direct patient care contact. In the process of assessing DIF, there are two common statistics used to determine if items are functioning differently (Linacre, 2012b). The first is the Mantel Chi-square (Mantel, 1963), which compares frequencies of respondents based on groups (i.e., cross-tabs of observations of the two groups) within each logit ability at the item level. In this analysis, a significant Chi-square value indicates association between group and ability, suggesting DIF. The second statistic is the Rasch-Welch t, which is a simple t-test comparison of the Rasch difficulty estimates between the groups based on a logistic regression model (Linacre,
2012b). Both statistics will be evaluated to determine if an item is functioning differently based on groupings. Linacre (2012b) notes that both statistics should produce the same results because they are based on an identical logit-linear theory of analysis, so both must provide statistically significant results to be concluded to have DIF based on the current analysis.

**DIF analysis of gender groups.** The first analysis sought to determine if items were functioning differently (i.e., more or less difficult) based on respondents’ gender. Since the sample was reduced to 222 respondents, 182 (81.98%) of them identified as female and 31 (13.96%) identified themselves as males. Nine respondents (4.05%) were excluded from the DIF analysis because their gender was not reported. The analysis procedure began with calculating relative item difficulties for the genders and determining the differences between them. It was hypothesized that the items would perform similarly across (i.e., there would be no difference between) genders.

All 30 items were analyzed and differences in item difficulties ranged from -1.36 to .77. Fourteen of the 30 differences were negative, suggesting those 14 items were relatively more difficult for females compared with 16 items being more difficult for males. Rasch-Welch t-values ranged from -1.96 to 1.60, indicating none of the differences were statistically significant (p > .05 for all). Mantel Chi-square values ranged from 2.569 to .000 (p > .05 for all), providing further support that the groups did not perform significantly different from each other on any of the items.

**DIF analysis of direct patient care contact groups.** The second analysis sought to determine if items were functioning differently (i.e., more or less difficult) based on
respondents’ role involving direct patient care contact. Of the 222 respondents, 160 (72.07%) of them identified as having direct patient care contact, while the remaining 62 (27.93%) did not report direct care contact. The analysis procedure mirrored that done previously to compare genders, where in this instance there was expected to be some items functioning differently based on respondents’ experience with direct patient care.

The differences in item difficulties ranged from -1.20 to 1.19, with 17 of the 30 items being easier, relatively, for those whose roles do not involve direct patient care contact (Table 28). The remaining 13 items were easier to endorse for participants who are involved in direct patient care contact. Rasch-Welch t-values ranged from -3.96 to 3.78, indicating the differences between groups on some items were statistically significant. Nine items (i.e., Items 1, 7, 10, 14, 15, 17, 18, 19, and 30) obtained statistically significant t-values (i.e., greater than 1.96, p < .05). Mantel Chi-square values ranged from 19.851 to .020, which also provided evidence of group differences. Ten items identified as displaying DIF (i.e., Items 1, 3, 6, 10, 14, 15, 17, 19, 21, and 30) achieved significant Mantel Chi-square values (p < .05) as well. Based on the results of both t and Chi-square values and their corresponding significance, Items 1, 10, 14, 15, 17, 19 and 30 are functioning differently between groups of respondents with and without direct patient care contact. Table 28 presents the DIF statistics for each item grouped by direct patient care contact.
Table 28.

*DIF Class Specification for Patient Care Contact Groups*

<table>
<thead>
<tr>
<th>Person Class</th>
<th>Obs-Exp Avg</th>
<th>DIF Measure</th>
<th>Person Class</th>
<th>Obs-Exp Avg</th>
<th>DIF Measure</th>
<th>DIF S.E.</th>
<th>DIF Contrast</th>
<th>Joint S.E.</th>
<th>Rasch-Welch t</th>
<th>df</th>
<th>Mantel Chi-Sq</th>
<th>Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.10</td>
<td>1.02</td>
<td>2</td>
<td>-0.26</td>
<td>1.93</td>
<td>0.18</td>
<td>-0.92</td>
<td>0.23</td>
<td>-3.96**</td>
<td></td>
<td>19.851**</td>
<td>Item 1</td>
</tr>
<tr>
<td>1</td>
<td>0.00</td>
<td>-0.20</td>
<td>2</td>
<td>0.01</td>
<td>-0.29</td>
<td>0.29</td>
<td>0.10</td>
<td>0.36</td>
<td>0.27</td>
<td></td>
<td>151</td>
<td>Item 2</td>
</tr>
<tr>
<td>1</td>
<td>0.03</td>
<td>0.16</td>
<td>2</td>
<td>-0.06</td>
<td>0.53</td>
<td>0.24</td>
<td>-0.36</td>
<td>0.30</td>
<td>-1.20</td>
<td></td>
<td>158</td>
<td>Item 3</td>
</tr>
<tr>
<td>1</td>
<td>0.03</td>
<td>0.71</td>
<td>2</td>
<td>-0.08</td>
<td>1.08</td>
<td>0.21</td>
<td>-0.37</td>
<td>0.27</td>
<td>-1.38</td>
<td></td>
<td>155</td>
<td>Item 4</td>
</tr>
<tr>
<td>1</td>
<td>0.00</td>
<td>-1.19</td>
<td>2</td>
<td>0.00</td>
<td>-1.19</td>
<td>0.39</td>
<td>0.00</td>
<td>0.49</td>
<td>0.00</td>
<td></td>
<td>157</td>
<td>Item 5</td>
</tr>
<tr>
<td>1</td>
<td>0.02</td>
<td>-2.38</td>
<td>2</td>
<td>-0.05</td>
<td>-1.18</td>
<td>0.39</td>
<td>-1.20</td>
<td>0.64</td>
<td>-1.88</td>
<td></td>
<td>196</td>
<td>Item 6</td>
</tr>
<tr>
<td>1</td>
<td>-0.05</td>
<td>0.54</td>
<td>2</td>
<td>0.12</td>
<td>-0.21</td>
<td>0.29</td>
<td>0.75</td>
<td>0.33</td>
<td>2.27*</td>
<td></td>
<td>136</td>
<td>Item 7</td>
</tr>
<tr>
<td>1</td>
<td>-0.04</td>
<td>0.78</td>
<td>2</td>
<td>0.11</td>
<td>0.23</td>
<td>0.25</td>
<td>0.56</td>
<td>0.30</td>
<td>1.86</td>
<td></td>
<td>139</td>
<td>Item 8</td>
</tr>
<tr>
<td>1</td>
<td>-0.02</td>
<td>1.02</td>
<td>2</td>
<td>0.06</td>
<td>0.74</td>
<td>0.23</td>
<td>0.28</td>
<td>0.27</td>
<td>1.03</td>
<td></td>
<td>143</td>
<td>Item 9</td>
</tr>
<tr>
<td>1</td>
<td>-0.04</td>
<td>0.33</td>
<td>2</td>
<td>0.11</td>
<td>-0.47</td>
<td>0.31</td>
<td>0.80</td>
<td>0.36</td>
<td>2.26*</td>
<td></td>
<td>135</td>
<td>Item 10</td>
</tr>
<tr>
<td>1</td>
<td>-0.02</td>
<td>0.68</td>
<td>2</td>
<td>0.05</td>
<td>0.41</td>
<td>0.24</td>
<td>0.27</td>
<td>0.29</td>
<td>0.92</td>
<td></td>
<td>144</td>
<td>Item 11</td>
</tr>
<tr>
<td>1</td>
<td>0.02</td>
<td>-0.15</td>
<td>2</td>
<td>-0.04</td>
<td>0.16</td>
<td>0.26</td>
<td>-0.31</td>
<td>0.33</td>
<td>-0.94</td>
<td></td>
<td>159</td>
<td>Item 12</td>
</tr>
<tr>
<td>1</td>
<td>0.00</td>
<td>-0.79</td>
<td>2</td>
<td>0.00</td>
<td>-0.79</td>
<td>0.34</td>
<td>0.00</td>
<td>0.43</td>
<td>0.00</td>
<td></td>
<td>155</td>
<td>Item 13</td>
</tr>
<tr>
<td>1</td>
<td>0.08</td>
<td>0.71</td>
<td>2</td>
<td>-0.20</td>
<td>1.53</td>
<td>0.19</td>
<td>-0.82</td>
<td>0.25</td>
<td>-3.25**</td>
<td></td>
<td>162</td>
<td>Item 14</td>
</tr>
<tr>
<td>1</td>
<td>-0.07</td>
<td>0.93</td>
<td>2</td>
<td>0.19</td>
<td>-0.13</td>
<td>0.28</td>
<td>1.06</td>
<td>0.32</td>
<td>3.33**</td>
<td></td>
<td>130</td>
<td>Item 15</td>
</tr>
<tr>
<td>1</td>
<td>0.01</td>
<td>-0.80</td>
<td>2</td>
<td>-0.03</td>
<td>-0.47</td>
<td>0.31</td>
<td>-0.33</td>
<td>0.40</td>
<td>-0.81</td>
<td></td>
<td>163</td>
<td>Item 16</td>
</tr>
<tr>
<td>1</td>
<td>0.10</td>
<td>1.31</td>
<td>2</td>
<td>-0.25</td>
<td>2.09</td>
<td>0.17</td>
<td>-0.78</td>
<td>0.22</td>
<td>-3.54**</td>
<td></td>
<td>159</td>
<td>Item 17</td>
</tr>
<tr>
<td>1</td>
<td>0.04</td>
<td>-0.67</td>
<td>2</td>
<td>-0.09</td>
<td>0.09</td>
<td>0.26</td>
<td>-0.76</td>
<td>0.36</td>
<td>-2.10*</td>
<td></td>
<td>172</td>
<td>Item 18</td>
</tr>
<tr>
<td>1</td>
<td>-0.03</td>
<td>-0.39</td>
<td>2</td>
<td>0.08</td>
<td>-1.34</td>
<td>0.41</td>
<td>0.95</td>
<td>0.47</td>
<td>2.03*</td>
<td></td>
<td>131</td>
<td>Item 19</td>
</tr>
<tr>
<td>1</td>
<td>-0.01</td>
<td>0.49</td>
<td>2</td>
<td>0.03</td>
<td>0.35</td>
<td>0.25</td>
<td>0.14</td>
<td>0.30</td>
<td>0.45</td>
<td></td>
<td>147</td>
<td>Item 20</td>
</tr>
</tbody>
</table>
Table 28. (continued)

<table>
<thead>
<tr>
<th>Person Class</th>
<th>Obs-Exp Avg</th>
<th>DIF Measure</th>
<th>Person Class</th>
<th>Obs-Exp Avg</th>
<th>DIF Measure</th>
<th>DIF S.E.</th>
<th>DIF Contrast S.E.</th>
<th>Joint Rasch-Welch t</th>
<th>Mantel Chi-Sq</th>
<th>Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.01</td>
<td>-1.11</td>
<td>2</td>
<td>0.04</td>
<td>-1.73</td>
<td>0.48</td>
<td>0.62</td>
<td>0.56</td>
<td>1.11</td>
<td>139</td>
</tr>
<tr>
<td>1</td>
<td>0.00</td>
<td>-0.80</td>
<td>2</td>
<td>-0.01</td>
<td>-0.67</td>
<td>0.33</td>
<td>-0.13</td>
<td>0.42</td>
<td>-0.30</td>
<td>158</td>
</tr>
<tr>
<td>1</td>
<td>-0.01</td>
<td>0.52</td>
<td>2</td>
<td>0.02</td>
<td>0.41</td>
<td>0.24</td>
<td>0.10</td>
<td>0.30</td>
<td>0.35</td>
<td>148</td>
</tr>
<tr>
<td>1</td>
<td>-0.01</td>
<td>-0.95</td>
<td>2</td>
<td>0.02</td>
<td>-1.18</td>
<td>0.39</td>
<td>0.24</td>
<td>0.48</td>
<td>0.49</td>
<td>149</td>
</tr>
<tr>
<td>1</td>
<td>0.00</td>
<td>-1.20</td>
<td>2</td>
<td>-0.01</td>
<td>-1.04</td>
<td>0.37</td>
<td>-0.16</td>
<td>0.48</td>
<td>-0.34</td>
<td>161</td>
</tr>
<tr>
<td>1</td>
<td>-0.02</td>
<td>1.71</td>
<td>2</td>
<td>0.06</td>
<td>1.53</td>
<td>0.19</td>
<td>0.18</td>
<td>0.23</td>
<td>0.79</td>
<td>143</td>
</tr>
<tr>
<td>1</td>
<td>0.02</td>
<td>-0.11</td>
<td>2</td>
<td>-0.05</td>
<td>0.23</td>
<td>0.25</td>
<td>-0.33</td>
<td>0.33</td>
<td>-1.02</td>
<td>159</td>
</tr>
<tr>
<td>1</td>
<td>0.01</td>
<td>-1.02</td>
<td>2</td>
<td>-0.02</td>
<td>-0.79</td>
<td>0.34</td>
<td>-0.24</td>
<td>0.44</td>
<td>-0.54</td>
<td>162</td>
</tr>
<tr>
<td>1</td>
<td>-0.02</td>
<td>-0.61</td>
<td>2</td>
<td>0.06</td>
<td>-1.34</td>
<td>0.41</td>
<td>0.73</td>
<td>0.48</td>
<td>1.53</td>
<td>136</td>
</tr>
<tr>
<td>1</td>
<td>-0.09</td>
<td>1.06</td>
<td>2</td>
<td>0.23</td>
<td>-0.13</td>
<td>0.28</td>
<td>1.19</td>
<td>0.32</td>
<td>3.78**</td>
<td>128</td>
</tr>
</tbody>
</table>

Note. Person class 1 refers to respondents with direct patient care contact, while person class 2 refers to respondents not engaging in direct patient care contact.

* *p < .05, ** *p < .01
The items identified as having DIF based on respondents’ direct patient care contact were Item 1 (i.e., “I am knowledgeable about patient care.”), Item 14 (i.e., “I have a strong impact on patient experience.”), and Item 30 (i.e., “I close every interaction by saying “Thank you” to patients and visitors.”). Based on the definition of DIF, the items were statistically different in terms of difficulty based on the ratings provided by the two groups. Examining the content of the items, this could have been predicted. Namely, those with direct patient contact found the first item, involving knowledge of patient care, much easier than employees without direct care contact. Further, employees noted they had a much stronger impact on patient experience when they were involved directly with patients, with those involved in direct patient contact finding the item much easier to endorse. The final item, closing interactions with a “Thank you”, was much easier to endorse for those not involved in direct patient care than for respondents who were. This makes sense based on the nature of interactions, where direct caregivers may naturally have different interactions with patients and/or visitors.

Together, the items perform similarly for males and females, where gender had no significant bearing on how difficult an item was to endorse. The experience of direct patient care contact, however, led for some items to be easier (i.e., being knowledgeable about patient care and having a strong impact on patient experience) while other items (i.e., closing every interaction with a “Thank you”) were more difficult to endorse, suggesting these behaviors are displayed less frequently. The identification of DIF in this situation is not necessarily problematic, as it could be if DIF was found between genders. Employees, as part of their roles, are engaged in different interactions with patients and
visitors and may find some behaviors easier or more difficult to perform consistently given the nature of those interactions.

The current research sought to evaluate the psychometric properties of the newly-developed service behavior instrument. Overall, the results suggest there were concerns based on participant responses, leading to questionable results regarding the psychometric properties. Results obtained in the pilot phase were interpreted with questionable validity based on the very limited sample obtained in the initial phase. Data were used to inform modifications to the scale; namely, the reduction of items to improve time-to-completion of the instrument. Data from the full-scale implementation provided evidence of universally high means for each of the items, suggesting each of the behaviors included in the instrument is frequently displayed by employees.

As a means of further data reduction following full-scale administration, a PCA was conducted to assist in identifying underlying factor structure within the items. The results of the PCA were confirmed using quasi-confirmatory techniques, which reflect a relationship among and grouping of items based on the Expected Service BehaviorSM from which they were developed. Together, a model consisting of 11 items comprising four factors was found to demonstrate statistical fit, likely through the reduction of measurement error by including non-significant parameters. These results provide supporting evidence that the items are inter-related and are moderately aligned with the behaviors from which they were developed. To investigate if the expected behaviors comprise different, unique dimensions within the measure, the items were analyzed using Rasch techniques.
A Rasch analysis was also conducted to evaluate the item fit with the construct, response category usage, dimensionality of the instrument, and potential for DIF. In this phase of analysis, the sample was further limited by the number of participants who recorded the maximum total score, representing the highest frequency of behavior for all of the items. Five items were identified as having poor fit with the Rasch model, but all were retained for the investigations into dimensionality and DIF. Based on the analysis of response category usage, the Andrich thresholds were disordered and the second choice was substantially under-used. The response categories of “Never or almost never” and “Rarely” were combined and analyzed again, providing support for a more clear and organized structure of response options. When person ability and item difficulty values were plotted against each other, it comes to support evidence provided in the item analysis that items were very easy to endorse for this population of respondents.

The dimensionality investigation sought to determine if the instrument is unidimensional (i.e., measuring only one construct), or if there is evidence of secondary factors based on the responses. The PCA of Rasch model residuals identified the potential for a second underlying factor with the strength of about two items (i.e., Eigenvalue of 2.4 in the first contrast). However, the loadings only exceeded .40 for five items, indicating there were few items to support two distinct dimensions. Further, once the items were separated into components, there was no clear distinction between the two based on content or target behavior. Together, the data indicate the presence of a single underlying dimension, representative of service quality.
The last analysis utilized the Rasch difficulty estimates to determine if items were functioning differently (i.e., more or less difficult) based on grouping of respondents. Two instances of group comparisons were made: first, difficulty estimates were compared across genders, and secondly between groups who reported direct patient care contact and those who did not. As anticipated, items functioned similarly between the genders, and no evidence of DIF was found using the gender comparisons. However, when direct patient care contact was used as the grouping variable, three items were identified as significantly differing in difficulty. These three items (i.e., Items 1, 14, and 30) were understandably different based on item content and the experience of direct patient care contact.

In summary, the CFA and Rasch analysis were used complimentary in the current research. The factor analytic methods were able to determine a number of underlying components, mainly aligning with the behaviors from which the items were generated. The results obtained in that phase remain unclear and were based on a limited number of items. These underlying components demonstrated a moderate to strong relationship with each other, as did the individual items, suggesting the existence of a single, fundamental factor. The later analysis was able to supplement that conclusion, where dimensionality testing provided little conclusive evidence of multiple dimensions present within the instrument. Though the results provide different interpretations of the item organization within behaviors, the more sound conclusions were provided by the Rasch investigation. Ultimately, the Rasch results provide support the items reflect one unidimensional construct: service quality.
CHAPTER V

DISCUSSION, LIMITATIONS, AND RECOMMENDATIONS

This chapter begins with a brief discussion and interpretation of the findings, expanding upon the information presented in the previous chapter. Following the interpretation, the implications for use within the current setting will be noted. Concluding this chapter will be a presentation of the limitations of the current investigation, as well as recommendations for future research. Subsequent investigations using this instrument and other thoughts regarding assessment for groups similar to the target population will be presented.

Discussion

The discussion section mirrors the methodology and results chapters, broken down into the phases of the investigation. Presented first is the pilot testing phase.

Pilot Testing Phase

The purpose of the pilot testing was to inform refinement and analysis of the full-scale instrument. The initial analysis plan for this phase included an item analysis and EFA to establish any underlying factor structure based on responses. Due to sample constraints, the validity of the latter analysis was deemed questionable. The item analysis provided useful information regarding item frequencies and distributions, leading to the conclusion that there were too many items included during this initial administration and questions were redundant when asked to provide ratings relevant to colleagues as well as patients and/or visitors. Thus, the pilot phase informed refinement through removal of the colleague category of items and removal of 16 items.
**Data collection and sample.** The pilot also provided information relative to the research process within this healthcare institution. Given survey requests were administered by the institution’s office dedicated to service quality (i.e., patient experience), the poor response rate was unforeseen. This was true even following some of the strategies for high response rates to online surveys mentioned by Nulty (2008), such as extending the duration of the survey’s availability and assuring anonymity of responses. The use of incentives, no matter how small, have been shown to substantially enhance survey completion percentage whether online or paper-and-pencil (Dommeyer, Baum, Hanna, & Chapman, 2004). Due to the nature of the healthcare setting, employees were not able to be compensated for their participation and thus were provided no incentive to participate. The lack of compensation may have negatively impacted the response rate of both pilot and full-scale administrations.

Unfortunately, the overall sample was not able to be expanded due to the limitations in place. The instrument was designed to be used with employees who had undergone training specific to the behaviors portrayed in the assessment. To be eligible to participate in the pilot phase, employees had to have received the training and have been within their first six months of employment. These stipulations were in place as the institution is currently in a roll-out phase of the training program and expanding the offerings beyond newly hired employees and those in need of service recovery. The sample remained limited to the individuals hired and trained by the institution, out of the control of the researcher. Employees provided the training as part of service recovery were included in the full-scale administration, which will be discussed later.
Analyses. The analyses conducted in the pilot phase were interpreted with caution based on the sample characteristics described above. The item analysis provided insight into participant responses, as employees self-reported a high frequency of all behaviors queried. Many items achieved exceptionally high means and were not able to distinguish between more and less quality-oriented employees. Further, the response frequencies indicated attrition of respondents based on the length of the survey and the number of responses they were asked to provide. The percent of items complete (Table 2) reflected participants’ willingness to begin, but less than half achieved more than 90% completion. The items displayed strong internal consistency, representing a level of similarity or agreement among the items. Together, the data suggested the final instrument should consist of fewer items, but include those that were most useful in discriminating among respondents.

The EFA was conducted though the sample size did not meet what many consider to be sufficient for this type of analysis (e.g., Thompson, 2004; Pett, Lackey, & Sullivan, 2003). Analysis progressed regardless, using all data available to provide evidence in support of instrument refinement. Two separate EFAs were conducted; the first included all 46 rating-scale items, while the second analyzed responses to 30 items following removal of items identified as problematic. The exploratory analyses provided evidence of 12 and nine factors, respectively, but no clear factor structure could be concluded based on item content. Ultimately, the EFA was able to identify statistical relationships among the items, but interpreting the groupings of the items did not yield a clear result. Such results were likely due to the constrained sample and limited variability within
responses, but provided information that was used during the instrument refinement phase.

Refinement. The analyses completed during the pilot phase provided insight into the items based on employee responses. As noted earlier, two major conclusions were brought out by the data obtained: 1) the instrument consists of redundant items or those not useful in measuring service quality, and 2) the amount of time to complete the survey is too long for this group of participants. The process of refinement is described in a previous section, so it will not be restated here. Ultimately, a critical review of the items and analyses was able to provide a reasonable solution to both issues – through removal of certain items and the redundancy of providing multiple ratings, a shorter version of the instrument was administered to the full-scale sample.

Full-Scale Administration

The refined measure was administered to a second cohort of employees at the same healthcare setting, using the same distribution methodology. The analyses sought to supplement the information provided in the pilot phase, further establishing validity and reliability of the new measure. A discussion of the results beginning with the data collection, sample, and analyses follows.

Data collection and sample. To supplement the limited sample provided in the pilot phase, a much larger cohort was selected in the full-scale administration. The instrument was sent to all employees who received the training, whether it was through new-hire orientation or as part of service recovery. The expansion to include employees in service recovery training provided a much larger sample for the full-scale
administration, and the response rate increased concomitantly. The corresponding rise of overall sample size and enhanced completion percent provided a response rate similar to, but still below, those discussed by Nunnaly (2008). While the response rate was more respectable than in the pilot phase, results may still be interpreted with caution as the appropriate sample size differed for each analysis used.

It cannot be overstated that generalizability of the conclusions should be limited to similar samples. The overall sample for the research, those who have been trained in patient care and high quality service using the healthcare institution’s expected behaviors, is a substantially narrow sample as the training program is still in its infancy. It is not outside the realm of possibility that other institutions or industries may expect the same behaviors out of their employees, but the instrument was developed using the single institution’s definition of high quality service. Thus, given the reasonable amount of time and frequency of training, the sample was intended to be all-inclusive. Limitations and concerns related to the population sampling and response rates will be discussed in a following section.

**Analyses.** The overarching purpose of the analyses using the full-scale administration data was to establish the psychometric properties of the instrument. The item analysis and the examination of underlying factors were re-hashed using the additional data collected in this phase in an effort to draw more reliable and valid conclusions. The factor structure was confirmed using quasi-confirmatory techniques, and a more detailed evaluation of dimensionality and item performance was completed using Rasch analysis. While the results indicate the factor structure was confirmed, that
conclusion was based on a small number of items and the inclusion of error covariances, which were non-significant parameters. The analysis is considered quasi-confirmatory as it does not adhere to the traditional standards of CFA by using the most parsimonious model as desired in the traditional CFA literature. Each will be dissected in separate sections below.

**Item analysis.** The results of the item analysis provide further support to the conclusions drawn in the pilot phase. Items continued to display high means, with values near the top of the range of the response scale. Internal consistency was also very strong, suggesting a strong relationship among items and an overall agreement in the construct being measured. The refined items were found to align more closely with one another, but concerns regarding the items’ ability to differentiate among respondents remained.

**Principal components analysis.** As noted previously, the EFA from the pilot phase did not provide theoretically sound conclusions and the results were interpreted with limited validity. Though a PCA was not part of the initial analysis plan, it was included as a means of data reduction prior to attempting CFA. Select items were removed from this portion of the analysis due to exceptionally high means, as well as items that were interpreted as “global” indicators of service quality or patient experience in nature. The results indicated sufficient alignment between 18 items based on the service behavior from which they were developed. Each component extracted from the data corresponded to a domain of the employee behavior indicative of good service quality mentioned in the literature or determined by the current institution. The PCA results provided a theoretical and conceptual basis on which to ground future research.
**Confirmatory factor analysis.** Several models were evaluated using confirmatory techniques, beginning with the model derived from the results of the PCA. This model consisted of the 18 items used in the previous analysis, and was not found to exhibit sufficient fit with the data. Model misfit was primarily a function of statistical redundancy among items (i.e., multicollinearity). Modifications to the original model included removal of several items to alleviate multicollinearity and the inclusion of error covariances, rooted in the method of data collection, to explain additional variance. This shifted the analysis for a pure confirmatory to a quasi-confirmatory approach in nature, as parameters were added to satisfy model fit. The final model was able to demonstrate statistical fit to the data, providing support for the structure of items aligning with the behaviors from which they were developed. This statistical fit is likely due to the elimination of measurement error using techniques not necessarily based on CFA theory. While the final model was able to meet the criteria for fit, the results were obtained using alternatives to the pure confirmatory approach.

**Rasch analysis.** Participants self-reported the frequency of their behaviors to be very high, supported by the Rasch analysis where there was little overlap between person ability and item difficulty estimates. The items overall displayed adequate fit with the Rasch model, but were interpreted as lacking difficulty based on participant responses. Similarly, the response categories reflecting infrequent behaviors were substantially under-used in relation to categories indicating higher frequencies of behaviors.

The results of the factor analysis of Rasch residuals provided evidence that the instrument assess one dimension, which represents that of “service quality.” There was
insufficient evidence and corresponding theory to determine a useful second dimension in the analysis. Namely, the conclusions from the confirmatory phase have been muted by the Rasch results due to the inclusive nature of the dimensionality investigation (i.e., the use of all 30 items) contrasted with the limited number of items included in the confirmatory analysis. The investigation into DIF determined that there were no differences among the items based on gender, but those with and without direct patient care contact saw a few items as significantly easier or more difficult based on their group membership. Together, the items combine to evaluate one dimension of overall service quality in agreement with the Rasch model. However, few items sufficiently discriminate among the respondents due to the high frequency with which employees display these behaviors. Furthermore, employees with and without direct patient care contact find some behaviors easier than others based on the nature of their interaction with patients and visitors.

Overall, the full-scale instrument was able to display some evidence of sound psychometric properties based on the analyses. There remain some areas to investigate for instrument improvement, but it can be concluded that service quality is accurately represented within the items. The items were developed using specific behavioral indicators presented by the healthcare institution, representing what they determined to be essential to quality service. Combining the items developed from the healthcare institution’s focal behaviors with items representative of global, overall service quality as evidenced in the existing literature, the items aligned with one another along a single dimension representing overall service quality. These conclusions are based mainly on
the Rasch results, where the CFA was able to include a much more limited number of items and still did not provide a clear interpretation for the instrument. The implications and limitations to these conclusions will be discussed in the following section.

The final instrument produced was slightly different from the pilot instrument. The full-scale instrument was refined to be a shorter, more precise version of the pilot instrument. Where the pilot instrument sought ratings from the perspectives of both patients/visitors and colleagues for all items, the full-scale instrument was more precise in the items for which patient/visitor and/or colleague ratings were requested. In this capacity, the pilot instrument was administered to identify which items best reflected employee behavior to be included on the full-scale instrument. Item wording and the response scale remained the same between the two instruments and the target behaviors were each represented within the full-scale instrument, as an attempt for the full-scale instrument to measure the same constructs as the pilot instrument.

Implications

The implications of the results seek to inform future research in the service quality domain, as well as the institutions utilizing the instrument. As it pertains to the service quality domain, the instrument developed in the current investigation was not able to demonstrate desirable psychometric qualities. The results suggest the instrument displays some evidence of reliability and validity, but results are inconclusive considering the limitations of the sample. The existing research on service quality has provided no evidence of an instrument evaluated using the Rasch method, so the current research sought to set a new standard for which instruments should strive to achieve.
Further, the strong inter-relationships among items and the evidence of a single underlying dimension indicate the behaviors determined by the healthcare institution are representative of high quality service. This became apparent as the newly-developed items performed similarly to those used in previous investigations of service quality, regardless of sector, as seen in the literature. The behaviors determined by the healthcare institution to be essential for high quality service were found to strongly relate with one another due the nature of each being important in service quality, reflective of the underlying dimension. This made the confirmatory analysis challenging, particularly as the individual items were expected to align together based on the behavior from which they were developed, but nonetheless shed light on the behaviors as deemed essential in high quality service.

What the results imply for the focal healthcare institution is that their employees are consistently exhibiting behaviors that are expected of them, based on self-report. Employees are frequently engaging with patients and visitors, showing active listening and empathy as they go about their duties. Institutions that aspire to utilize the instrument may benefit from the psychometric properties established based on the results. Data suggest the items align with a single dimension, representing service quality, to which institutions could evaluate their employees. The caveat to this being that while the items align with one healthcare institution’s definition of service quality based on expected behaviors of employees, this may not be applicable to all healthcare settings or other industries.
Limitations

The first major limitation to the current research is based on the sample. The target population was defined a priori, as the instrument was developed to evaluate employees who had undergone training on service behaviors. Employees were granted participation only as new-hires or as part of service recovery, so the number of people receiving the training was small relative to the number of employees within the healthcare institution. The captive sample could only be grown through the hiring and service recovery identification process, so a larger sample was challenging to secure. The number of completed instruments in both the pilot and full-scale phases were well below expectation, rarely meeting the sample size requirements for the conducted analyses. All analyses should be interpreted with caution based on the procedures used to achieve the sought-after sample sizes for each analysis.

Specifically, there were concerns in acquiring an adequate sample to conduct the EFA, PCA, CFA, and Rasch analyses. To rectify the initial concerns for the later analyses, the results from the pilot sample were included with those from the full sample beginning with the PCA. A subsample of 150 respondents was selected from the combined sample to satisfy the criteria of the 5:1 persons to items ratio for exploratory analyses. The full sample of respondents was used in the confirmatory analysis, which could have impacted the findings. The use of PCA and the procedure for data reduction is met with debate in the social science literature (see Thompson, 2004), questioning the appropriateness of the method. The procedure was followed here as it was outlined in previously published studies incorporating EFA, PCA, and CFA using the same data.
(Matsunaga, 2010). As for combining the samples, the full-scale data was seen as an extension of the pilot instrument, so questions that were asked in both were paired based on time being the only confounding factor between administrations.

The confirmatory analysis was considered quasi-confirmatory as the analysis method shifted based on the limitations within the data. In order to account for the inter-relationships among items, error covariances were included and retained even when they did not meet the criteria for significance. Modifications to the model were made in order to generate adequate fit between the data and the model, without theory or precedent as guidance. Items were reduced and few factors were retained to develop a model that fit the data adequately. The latent variables that emerged may be subject to questionable reliability and validity, again due to limitations within the data.

Secondly, there were limitations to the research based on the secondary nature of the data. The instrument distribution was overseen by the healthcare institution, not the researcher. The concerns with this approach are mixed. While responses may have been influenced had an external researcher requested participation, the participants could be just as likely to be responsive given the participation request was coming from their corporate office. There were no indications the instrument was more or less likely to be completed given it was distributed by inter-corporation email, but should be considered in future studies of similar scope.

There are also limitations evident within the instruments. The pilot instrument was developed to be comprehensive and all-inclusive given the review of literature, criteria set by the healthcare institution, and the team of experts involved in item
development. The process of identifying problematic items yielded a larger number of items than preliminarily thought. Even after critically reviewing and refining the measure, a factor structure consisting of only 11 items could be confirmed allowing non-significant error covariance terms to remain included in the model. This could be due to the nature of the items and their inter-relatedness within the construct, or could be the result of the available response options. Given the questions probed frequency of behaviors, the scale ranging from “Always or almost always” to “Never or almost never” did not show enough sensitivity to appropriately distinguish among respondents to the extent anticipated.

Lastly, the use of self-report as a research method is always subject to honesty in participant report. Though employees were asked to provide input as to how they think patients, visitors, and colleagues would rate their frequency of behavior, there was no method to triangulate the findings. Thus, the only values for each employee were their own self-reported behavioral frequencies. Anonymity of responses and no connection with employee records were the methods in place to control for biased responses, but the extent to which responses were impacted is unknown.

**Recommendations**

Based on the limitations of the current research, some suggestions for future research include a strict approach to refinement, triangulation of self-report data, the use of paper-and-pencil survey methodology, and expanding the research to include participants not included in the training. All avenues should be considered in developing a new service behavior instrument or further testing of the current instrument.
First of all, the refinement phase of the current investigation provided insight into future instruments assessing service quality. This phase of the research suggested the instrument should be brief and succinct, collecting data that are representative of service quality within the unique industry. The researcher should approach the instrument development with clarity, avoiding redundancy when possible. Similarly, instruments that are too long present challenges to respondents based on the amount of time to complete and potential fatigue, even though data are self-report.

Secondly, the results provided by the employee self-report could be further validated through the use of similar measures provided to supervisors, colleagues, and patients. The other sources of information could serve to validate the employee self-report so long as patients, visitors, and colleagues report employee behaviors at a similar frequency to which employees do themselves. Given the nature of this instrument, it could be adapted easily for distribution to these other stakeholder groups, collecting valuable data in triangulating the findings. Supervisors, colleagues, and patients or visitors who have interacted with the employee should be able to provide accurate ratings for most of the items, providing a theoretically unbiased evaluation of employee behavior with minimal change to the existing instrument.

In the current investigation, respondents may have provided inflated ratings of these behaviors due to the self-report nature of the data. That can be overcome in future research by employing the triangulation techniques of including other raters (i.e., patients, colleagues, or supervisors as discussed above), or inclusion of other methods (Creswell & Plano Clark, 2010). Namely, a direct observation would provide meaningful
data as the behaviors are overt and able to be viewed by others. Participant ratings would then be subject to actual observation of these behaviors, as opposed to potentially-biased self-reported frequencies. Nonetheless, participant self-report of these desired behaviors should be thoroughly considered in future investigations, where data need to be validated using additional methods and/or sources of data.

Additionally, an alternative to the online survey method should be considered. Due to the nature of the sample being employees who have participated in the training, a better response rate could be achieved if participants were given a paper-and-pencil test as opposed to an electronic version. The training could be adjusted to allow for a brief meeting after several weeks, at which point the instrument could be administered. Other research has found evidence that response rate improves when using a paper-based assessment (Dommeyer et al., 2004), so providing a set time for employees to complete would substantially improve the number of responses. Regardless of survey mode in future administrations, the use of some incentives should be considered when applicable.

Future research should investigate employees who were not participants of the training. Including employees without this explicit training may gather a more realistic portrayal of service quality without the impact of the training. In this way, the instrument can serve as a baseline evaluation of employee service quality to allow comparisons before and after training or among employees in groups that receive training versus those who did not receive the training. This information would further serve to validate the measure beyond what was established in this investigation. In its current state, this
instrument could serve as a criterion-related measure for good service quality, providing a level to which healthcare institutions should strive to achieve.

The nature of the healthcare industry should be considered throughout future investigations. Namely, the methodological approach of sampling healthcare workers is challenging based on the nature of their work, consistently navigating between patients and multi-tasking in a fast-paced environment. Methodological concerns may be overcome with a more practical approach to the healthcare setting, which may include the paper-and-pencil method noted above, a live researcher collecting data, or an extended timeline for data collection. Given the unique demands placed on healthcare workers, there should be special consideration to how data are collected, which may have been overestimated as part of the current investigation’s methodology.

The most salient recommendation lies in the healthcare institution’s value of collecting data. Future research in this area needs to be deemed valuable by institution leadership. During the early stages of the current investigation, the institution’s leadership was very motivated to collect the information, which diminished as leadership recognized employees were not eager and willing to participate without incentives. Researchers and leaders within the institution need to emphasize the importance of the data that is collected as it informs strengths and weaknesses of the institution’s personnel. Future research in this area relies heavily upon a strong working relationship between researchers and the institution.

Taken together, the literature has outlined an established need for measuring employee behaviors as they relate to patient satisfaction. This reliance on patient
satisfaction presents a trend in healthcare policy that is currently under way. As the paradigm shifts toward emphasizing patient satisfaction, the industry has a tremendous need for a reliable, valid instrument assessing patient satisfaction from within. This instrument was intended to meet the unique needs of the healthcare industry, providing institutions the ability to evaluate their needs, impacts, and connect their focus of employees with the changing priorities of the industry. While this instrument did not meet the strict criteria set out for establishing reliability and validity, the current investigation provides helpful insight for future research in this area.

The development of this instrument incorporates many distinctive approaches compared with previous instruments evaluating service quality. Few existing instruments have examined service quality through pre-defined dimensions, which are represented here by the behavioral indicators identified by the institution. Unique to the healthcare setting, this instrument presents a novel approach of collecting employee self-report data, allowing the institution to identify weaknesses in specific areas of service quality. This is in stark contrast to the healthcare status quo, where patients have been asked to provide ratings of the staff as a whole. Finally, other existing service quality instruments have yet to be subject to advanced psychometric analyses, as was the case of this investigation, employing confirmatory factor and Rasch analyses.

This instrument was developed with the intention of meeting the unique needs of the healthcare industry. Based on the results of the current investigation, the newly-revised instrument can be used to shape the path of assessing service quality in the healthcare setting in the future. This research represents only the beginning of
establishing an instrument at the forefront of patient satisfaction and service quality, as this instrument will be subject to more revision, rigorous testing, and further implementation. Additional data collection and analysis methods, including incorporating a longitudinal approach, would provide useful insight into drawing concrete conclusions about the utility of this instrument beyond the current investigation.
APPENDICES
APPENDIX A

PILOT INSTRUMENT
Appendix A

Pilot Instrument

Do you have direct patient care contact?

- Yes
- No

Please complete each of the following items honestly by providing an indication of how you anticipate patients/visitors and colleagues will rate you in your role as a Cleveland Clinic caregiver.

Rate yourself by selecting the most appropriate option from each drop down box.

<table>
<thead>
<tr>
<th>Patients/Visitors would say that...</th>
<th>Colleagues would say that...</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. I am knowledgeable about patient care.</td>
<td>▼</td>
</tr>
<tr>
<td>2. I acknowledge visitors and welcome them.</td>
<td>▼</td>
</tr>
<tr>
<td>3. I take patients/visitors’ complaints personally.</td>
<td>▼</td>
</tr>
<tr>
<td>4. I introduce myself when I meet patients/visitors.</td>
<td>▼</td>
</tr>
<tr>
<td>5. I make patients/visitors feel comfortable.</td>
<td>▼</td>
</tr>
<tr>
<td>6. I use patients/visitors’ preferred name when I speak with them.</td>
<td>▼</td>
</tr>
<tr>
<td>7. If I’m having a challenging day, patients/visitors can tell.</td>
<td>▼</td>
</tr>
<tr>
<td>8. I am genuine when interacting with patients/visitors.</td>
<td>▼</td>
</tr>
</tbody>
</table>
Please complete each of the following items honestly by providing an indication of how you anticipate patients/visitors and colleagues will rate you in your role as a Cleveland Clinic caregiver.

Rate yourself by selecting the most appropriate option from each drop down box.

<table>
<thead>
<tr>
<th></th>
<th>Patients/Visitors would say that...</th>
<th>Colleagues would say that...</th>
</tr>
</thead>
<tbody>
<tr>
<td>9) I treat patients with care.</td>
<td>[ ] ▼</td>
<td>[ ] ▼</td>
</tr>
<tr>
<td>10) Each interaction I have with patients/visitors is authentic and individualized.</td>
<td>[ ] ▼</td>
<td>[ ] ▼</td>
</tr>
<tr>
<td>11) I summarize back to patients/visitors what they communicate with me.</td>
<td>[ ] ▼</td>
<td>[ ] ▼</td>
</tr>
<tr>
<td>12) I can follow through on my responsibilities and still meet patients’ expectations.</td>
<td>[ ] ▼</td>
<td>[ ] ▼</td>
</tr>
<tr>
<td>13) I encourage patients/visitors to discuss or present any concerns throughout their visit.</td>
<td>[ ] ▼</td>
<td>[ ] ▼</td>
</tr>
<tr>
<td>14) “Polite” and “thoughtful” are two words patients/visitors use to describe me.</td>
<td>[ ] ▼</td>
<td>[ ] ▼</td>
</tr>
<tr>
<td>15) I anticipate the needs of my patients/visitors.</td>
<td>[ ] ▼</td>
<td>[ ] ▼</td>
</tr>
<tr>
<td>16) I clearly communicate to patients/visitors what they can expect from me.</td>
<td>[ ] ▼</td>
<td>[ ] ▼</td>
</tr>
</tbody>
</table>
Please complete each of the following items honestly by providing an indication of how you anticipate patients/visitors and colleagues will rate you in your role as a Cleveland Clinic caregiver.

Rate yourself by selecting the most appropriate option from each drop down box.

<table>
<thead>
<tr>
<th>Item</th>
<th>Patients/Visitors would say that...</th>
<th>Colleagues would say that...</th>
</tr>
</thead>
<tbody>
<tr>
<td>17.</td>
<td>I contact the appropriate person to address patients/visitors’ concerns.</td>
<td>▼</td>
</tr>
<tr>
<td>18.</td>
<td>I demonstrate an interest in patients/visitors and their needs.</td>
<td>▼</td>
</tr>
<tr>
<td>19.</td>
<td>I share with patients/visitors resources that can help them resolve their concerns.</td>
<td>▼</td>
</tr>
<tr>
<td>20.</td>
<td>I focus on patients/visitors when they communicate with me.</td>
<td>▼</td>
</tr>
<tr>
<td>21.</td>
<td>I have a strong impact on patient experience.</td>
<td>▼</td>
</tr>
<tr>
<td>22.</td>
<td>I provide patients/visitors with feedback (verbal or nonverbal) when they are talking with me.</td>
<td>▼</td>
</tr>
<tr>
<td>23.</td>
<td>I am considerate to patients/visitors.</td>
<td>▼</td>
</tr>
<tr>
<td>24.</td>
<td>I work collaboratively with my colleagues.</td>
<td>▼</td>
</tr>
</tbody>
</table>

Please complete each of the following items honestly by providing an indication of how you anticipate patients/visitors and colleagues will rate you in your role as a Cleveland Clinic caregiver.

Rate yourself by selecting the most appropriate option from each drop down box.

<table>
<thead>
<tr>
<th>Item</th>
<th>Patients/Visitors would say that...</th>
<th>Colleagues would say that...</th>
</tr>
</thead>
<tbody>
<tr>
<td>25.</td>
<td>I am forthcoming in sharing what patients can expect from any procedure.</td>
<td>▼</td>
</tr>
<tr>
<td>26.</td>
<td>I am prompt in my service delivery.</td>
<td>▼</td>
</tr>
<tr>
<td>27.</td>
<td>I seek to build a positive, helpful relationship with patients/visitors.</td>
<td>▼</td>
</tr>
<tr>
<td>28.</td>
<td>I encourage patients/visitors to ask questions.</td>
<td>▼</td>
</tr>
<tr>
<td>29.</td>
<td>I ask others how they prefer to address them (i.e., what they like to be called).</td>
<td>▼</td>
</tr>
<tr>
<td>30.</td>
<td>I am timely in my response to patients/visitors’ concerns or requests.</td>
<td>▼</td>
</tr>
<tr>
<td>31.</td>
<td>I make eye contact when speaking with patients/visitors.</td>
<td>▼</td>
</tr>
<tr>
<td>32.</td>
<td>I explain my title or role.</td>
<td>▼</td>
</tr>
</tbody>
</table>
Please complete each of the following items honestly by providing an indication of how you anticipate patients/visitors and colleagues will rate you in your role as a Cleveland Clinic caregiver.

Rate yourself by selecting the most appropriate option from each drop down box.

<table>
<thead>
<tr>
<th>Patients/Visitors would say that...</th>
<th>Colleagues would say that...</th>
</tr>
</thead>
<tbody>
<tr>
<td>33) ... I greet others warmly.</td>
<td>▼</td>
</tr>
<tr>
<td>34) ... I do my best to assure patients/visitors' needs are met.</td>
<td>▼</td>
</tr>
<tr>
<td>35) ... I am mindful of patients' privacy and dignity.</td>
<td>▼</td>
</tr>
<tr>
<td>36) ... I respond with appropriate emotions when interacting with patients/visitors.</td>
<td>▼</td>
</tr>
<tr>
<td>37) ... I am sincere in my apologies.</td>
<td>▼</td>
</tr>
<tr>
<td>38) ... I acknowledge and respect patients/visitors' emotions.</td>
<td>▼</td>
</tr>
<tr>
<td>39) ... I do my best to show compassion for patients or families.</td>
<td>▼</td>
</tr>
<tr>
<td>40) ... I encourage patients/visitors to ask for help when necessary.</td>
<td>▼</td>
</tr>
</tbody>
</table>

Please complete each of the following items honestly by providing an indication of how you anticipate patients/visitors and colleagues will rate you in your role as a Cleveland Clinic caregiver.

Rate yourself by selecting the most appropriate option from each drop down box.

<table>
<thead>
<tr>
<th>Patients/Visitors would say that...</th>
<th>Colleagues would say that...</th>
</tr>
</thead>
<tbody>
<tr>
<td>41) ... When I forward a patient/visitor's concern, I follow up to ensure it has been handled.</td>
<td>▼</td>
</tr>
<tr>
<td>42) ... When I apologize, I take care not to place blame.</td>
<td>▼</td>
</tr>
<tr>
<td>43) ... I can clearly or explain what patients/visitors do not understand.</td>
<td>▼</td>
</tr>
<tr>
<td>44) ... I strive to be professional in my interactions.</td>
<td>▼</td>
</tr>
<tr>
<td>45) ... I recognize that each patient/visitor has a unique experience and circumstances.</td>
<td>▼</td>
</tr>
<tr>
<td>46) ... I close every interaction by saying “Thank you” to patients/visitors.</td>
<td>▼</td>
</tr>
</tbody>
</table>
Please read the following scenarios carefully and select the answer that best describes how you would respond in each situation.

1) As you enter the building to begin your shift, you are asked by a visitor to help him locate the nearest ATM. You do not have a lot of time before you need to report for your shift, and the nearest ATM is rather far away. What do you do?
   a) Apologize to the visitor and report to work.
   b) Walk the visitor to the ATM and then report report to work late.
   c) Explain you are about to report to work and ask the visitor to ask someone else.
   d) Give the visitor detailed directions on where the ATM is and report to work.

2) A patient is in a lengthy and invasive surgery, and she has visitors waiting in the lobby as you are passing through. Though you have not met them before, the visitors are visibly nervous and have stopped you to ask what is going on. What would you say in response?
   a) “I’m sorry yours feeling that way, but I do not know.”
   b) “I understand you might be nervous. Let me try to find someone.”
   c) “I know you’re upset. Your doctor and nurse will let you know as soon as possible.”
   d) “I understand your frustration and anxiety. I’m sure someone will be with you shortly.”

3) A patient has come in for a procedure, and she seems nervous. You need to perform a few tasks before moving on. What do you do?
   a) Share that it’s okay to be nervous and that many patients are before procedures.
   b) Explain to her what you are doing.
   c) Ask for clarification of what her concerns are.
   d) Perform your tasks in an efficient manner.

4) You enter the room to complete an assigned task related to your job, and the patient appears to be sleeping. You do not need to wake him to do the task. He has visitors present who are watching television in the room. What do you do?
   a) Greet the visitors with eye contact and a smile, making sure to let the patient sleep.
   b) Work quietly and quietly, taking care not to disturb the patient or his visitors.
   c) Introduce yourself to the visitors and include your role.
   d) Make yourself visible to the visitors and ask if there’s anything you can do for them or the patient.
6) You are performing some simple tasks in a patient’s room. While your tasks do not directly involve her, she asks a few questions. What do you do?

- a) Stop what you are doing, turn to the patient, and answer her questions.
- b) Participate in the conversation while performing your tasks.
- c) Explain that you can answer her questions as soon as you are done with your task.
- d) Explain to her what you are doing and provide her with information on your task.

6) A patient and his visitors appear very nervous as they await further information on test results. You do not have the information they want to know. What would you say in response?

- a) “I understand this might be stressful for you. You will find out the results as soon as possible.”
- b) “I’m sure everything will be fine. Is there something I can do for you while you wait?”
- c) “Some of the results may take a while, depending on the test. It’s worth the wait.”
- d) “I’m sorry it is taking so long, but there are a lot of reasons test results get delayed.”

7) You are working with a patient who is very difficult during interactions. You have tried several times to establish a positive relationship with her, but she is not reciprocating. What do you do?

- a) Accept that there is not anything you can do if the patient does not reciprocate with you.
- b) Be aware that she is in a difficult situation and continue to be friendly.
- c) Share with her your perceptions and explain you are trying to build rapport with her.
- d) Explain your professional responsibilities to her in a way she may understand.

8) When you enter the room, the patient and his visitors change their tone of voice so they are now whispering. You aren’t quite sure what they are talking about. What do you do?

- a) Ask them politely if you can be of help to them.
- b) Disregard their whispering because they are not speaking with you and go about completing your tasks.
- c) Ask if they would like privacy while having their discussion.
- d) Smile and make eye contact while you go about completing your tasks.
APPENDIX B

FULL-SCALE INSTRUMENT
Appendix B

Full-Scale Instrument

Do you have direct patient care contact?

- Yes
- No

Please complete each of the following items honestly by providing an indication of how you anticipate patients/visitors will rate you in your role as a Cleveland Clinic caregiver.

Rate yourself by selecting the most appropriate option for each item.

<table>
<thead>
<tr>
<th>Patients/Visitors would say that...</th>
<th>1 – Never or Almost Never</th>
<th>2 – Rarely</th>
<th>3 – Sometimes</th>
<th>4 – Frequently</th>
<th>5 – Always or Almost Always</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) I am knowledgeable about patient care.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2) I acknowledge visitors and welcome them.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3) I introduce myself when I meet patients/visitors.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4) I use patients/visitors’ preferred name when I speak with them.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5) I am genuine when interacting with patients/visitors.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6) I treat patients with care.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7) I contact the appropriate person to address patients/visitors’ concerns.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8) I can follow through on my responsibilities and still meet patients’ expectations.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9) I encourage patients/visitors to discuss or present any concerns throughout their visit.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10) “Paite” and “thoughtful” are two words patients/visitors use to describe me.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11) I clearly communicate to patients/visitors what they can expect from me.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12) I demonstrate an interest in patients/visitors and their needs.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patients/Visitors would say that...</td>
<td>1 – Never or Almost Never</td>
<td>2 – Rarely</td>
<td>3 – Sometimes</td>
<td>4 – Frequently</td>
<td>5 – Always or Almost Always</td>
</tr>
<tr>
<td>-----------------------------------</td>
<td>---------------------------</td>
<td>------------</td>
<td>--------------</td>
<td>---------------</td>
<td>---------------------------</td>
</tr>
<tr>
<td>13) ...I focus on patients/visitors when they communicate with me.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14) ...I have a strong impact on patient experience.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15) ...I am prompt in my service delivery.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16) ...I work collaboratively with my colleagues.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17) ...I am forthcoming in sharing what patients can expect from any procedure.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18) ...I seek to build a positive, helping relationship with patients/visitors.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Patients/Visitors would say that...</th>
<th>1 – Never or Almost Never</th>
<th>2 – Rarely</th>
<th>3 – Sometimes</th>
<th>4 – Frequently</th>
<th>5 – Always or Almost Always</th>
</tr>
</thead>
<tbody>
<tr>
<td>19) ...I make eye contact when speaking with patients/visitors.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20) ...I explain my title or role.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21) ...I greet others warmly.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22) ...I do my best to assure patients/visitors' needs are met.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23) ...I encourage patients/visitors to ask questions.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24) ...I am sincere in my apologies.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Patients/Visitors would say that...</th>
<th>1 – Never or Almost Never</th>
<th>2 – Rarely</th>
<th>3 – Sometimes</th>
<th>4 – Frequently</th>
<th>5 – Always or Almost Always</th>
</tr>
</thead>
<tbody>
<tr>
<td>25) ...I do my best to show compassion for patients or families.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26) ...When I respond to a patient's concern, I follow up to ensure it has been handled.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27) ...I encourage patients/visitors to ask for help when necessary.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28) ...I strive to be professional in my interactions.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29) ...I recognize that each patient/visitor has a unique experience and circumstances.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30) ...I close every interaction by saying “Thank you” to patients/visitors.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Please read each of the following 7 scenarios carefully and select the answer that best describes how you would respond in each situation.

1) As you enter the building to begin your shift, you are asked by a visitor to help him locate the nearest ATM. You do not have a lot of time before you need to report for your shift, and the nearest ATM is rather far away. What do you do?

   a) Apologize to the visitor and report to work.
   b) Walk the visitor to the ATM and then report report to work late.
   c) Explain you are about to report to work and ask the visitor to ask someone else.
   d) Give the visitor detailed directions on where the ATM is and report to work.

2) A patient has come in for a procedure, and she seems nervous. You need to perform a few tasks before moving on. What do you do?

   a) Share that it’s okay to be nervous and that many patients are before procedures.
   b) Explain to her what you are doing.
   c) Ask for clarification of whether concerns are.
   d) Perform your tasks in an efficient manner.

3) You enter the room to complete an assigned task related to your job, and the patient appears to be sleeping. You do not need to wake him to do the task. He has visitors present who are watching television in the room. What do you do?

   a) Greet the visitors with eye contact and a smile, making sure to let the patient sleep.
   b) Work quickly and quietly, taking care not to disturb the patient or his visitors.
   c) Introduce yourself to the visitors and include your role.
   d) Make yourself visible to the visitors and ask if there’s anything you can do for them or the patient.

4) You are performing some simple tasks in a patient’s room. While your tasks do not directly involve her, she asks a few questions. What do you do?

   a) Stop what you are doing, turn to the patient, and answer her questions.
   b) Participate in the conversation while performing your tasks.
   c) Explain that you can answer her questions as soon as you are done with your task.
   d) Explain to her what you are doing and provide her with information on your task.
5) A patient and his visitors appear very nervous as they await further information on test results. You do not have the information they want to know. What would you say in response?

a) "I understand this might be stressful for you. I will find out the results as soon as possible."

b) "I'm sure everything will be fine. Is there something I can do for you while you wait?"

c) "Some of the results may take a while, depending on the test. It's worth the wait."

d) "I'm sorry it is taking so long, but there are a lot of reasons test results get delayed."

6) When you enter the room, the patient and his visitors change their tone of voice so they are now whispering. You aren't quite sure what they are talking about. What do you do?

a) Ask them politely if you can be of help to them.

b) Disregard their whispering because they are not speaking with you and go about completing your tasks.

c) Ask if they would like privacy while having their discussion.

d) Smile and make eye contact while you go about completing your tasks.

7) You are working with a patient who is very difficult during interactions. You have tried several times to establish a positive relationship with her, but she is not reciprocating. What do you do?

a) Accept that there isn’t anything you can do if the patient does not reciprocate with you.

b) Share with her your perceptions and explain you are trying to build rapport with her.

c) Explain your professional responsibilities to her in a way she may understand.

d) Be aware that she is in a difficult situation and continue to be friendly.

Please complete each of the following items honestly by providing an indicator of how you anticipate your colleagues would rate you in your role as a Cleveland Clinic caregiver.

Rate yourself by selecting the most appropriate option for each item.

<table>
<thead>
<tr>
<th>My Colleagues would say that...</th>
<th>1 – Never or Almost Never</th>
<th>2 – Rarely</th>
<th>3 – Sometimes</th>
<th>4 – Frequently</th>
<th>5 – Always or Almost Always</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) I treat patients with care.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2) I am considerate to patients/visitors.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3) I focus on patients/visitors when they communicate with me.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4) If I'm having a challenging day, patients/visitors can tell.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5) I am prompt in my service delivery.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6) Each interaction I have with patients/visitors is authentic and individualized.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7) I anticipate the needs of my patients/visitors.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8) I strive to be professional in my interactions.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9) I am mindful of patients' privacy and dignity.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX C

RHO CORRELATIONS BETWEEN PATIENT/VISITOR AND COLLEAGUE RATINGS
## Appendix C

### Rho Correlations Between Patient/Visitor and Colleague Ratings

<table>
<thead>
<tr>
<th>#</th>
<th>Rho</th>
<th>Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.878</td>
<td>I am knowledgeable about patient care.</td>
</tr>
<tr>
<td>2</td>
<td>0.915</td>
<td>I acknowledge visitors and welcome them.</td>
</tr>
<tr>
<td>3</td>
<td>0.963</td>
<td>I take patients'/visitors' complaints personally.</td>
</tr>
<tr>
<td>4</td>
<td>0.976</td>
<td>I introduce myself when I meet patients/visitors.</td>
</tr>
<tr>
<td>5</td>
<td>0.836</td>
<td>I make patients/visitors feel comfortable.</td>
</tr>
<tr>
<td>6</td>
<td>0.964</td>
<td>I use patients'/visitors' preferred name when I speak with them.</td>
</tr>
<tr>
<td>7</td>
<td>0.823</td>
<td>If I'm having a challenging day, patients/visitors can tell.</td>
</tr>
<tr>
<td>8</td>
<td>0.867</td>
<td>I am genuine when interacting with patients/visitors.</td>
</tr>
<tr>
<td>9</td>
<td>0.499</td>
<td>I treat patients with care.</td>
</tr>
<tr>
<td>10</td>
<td>0.945</td>
<td>Each interaction I have with patients/visitors is authentic and individualized.</td>
</tr>
<tr>
<td>11</td>
<td>0.907</td>
<td>I summarize back to patients/visitors what they communicate with me.</td>
</tr>
<tr>
<td>12</td>
<td>0.941</td>
<td>I can follow through on my responsibilities and still meet patients' expectations.</td>
</tr>
<tr>
<td>13</td>
<td>0.965</td>
<td>I encourage patients/visitors to discuss or present any concerns throughout their visit.</td>
</tr>
<tr>
<td>14</td>
<td>0.915</td>
<td>&quot;Polite&quot; and &quot;thoughtful&quot; are two words patients/visitors use to describe me.</td>
</tr>
<tr>
<td>15</td>
<td>0.912</td>
<td>I anticipate the needs of my patients/visitors.</td>
</tr>
<tr>
<td>16</td>
<td>0.951</td>
<td>I clearly communicate to patients/visitors what they can expect from me.</td>
</tr>
<tr>
<td>17</td>
<td>0.919</td>
<td>I contact the appropriate person to address patients'/visitors' concerns.</td>
</tr>
<tr>
<td>18</td>
<td>0.827</td>
<td>I demonstrate an interest in patients/visitors and their needs.</td>
</tr>
<tr>
<td>19</td>
<td>0.918</td>
<td>I share with patients/visitors resources that can help them resolve their concerns.</td>
</tr>
<tr>
<td>20</td>
<td>0.826</td>
<td>I focus on patients/visitors when they communicate with me.</td>
</tr>
<tr>
<td>21</td>
<td>0.945</td>
<td>I have a strong impact on patient experience.</td>
</tr>
<tr>
<td>22</td>
<td>0.915</td>
<td>I provide patients/visitors with feedback (verbal or nonverbal) when they are talking with me.</td>
</tr>
<tr>
<td>23</td>
<td>0.789</td>
<td>I am considerate to patients/visitors.</td>
</tr>
<tr>
<td>24</td>
<td>0.848</td>
<td>I work collaboratively with my colleagues.</td>
</tr>
<tr>
<td>25</td>
<td>0.940</td>
<td>I am forthcoming in sharing what patients can expect from any procedure.</td>
</tr>
<tr>
<td>26</td>
<td>0.924</td>
<td>I am prompt in my service delivery.</td>
</tr>
<tr>
<td>27</td>
<td>0.872</td>
<td>I seek to build a positive, helping relationship with patients/visitors.</td>
</tr>
<tr>
<td>28</td>
<td>0.912</td>
<td>I encourage patients/visitors to ask questions.</td>
</tr>
<tr>
<td>29</td>
<td>0.973</td>
<td>I ask others how they prefer I address them (i.e. what they like to be called).</td>
</tr>
<tr>
<td>#</td>
<td>Rho</td>
<td>Item</td>
</tr>
<tr>
<td>----</td>
<td>------</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>30</td>
<td>0.945</td>
<td>I am timely in my response to patients'/visitors' concerns or requests.</td>
</tr>
<tr>
<td>31</td>
<td>1.000</td>
<td>I make eye contact when speaking with patients/visitors.</td>
</tr>
<tr>
<td>32</td>
<td>0.909</td>
<td>I explain my title or role.</td>
</tr>
<tr>
<td>33</td>
<td>0.861</td>
<td>I greet others warmly.</td>
</tr>
<tr>
<td>34</td>
<td>0.944</td>
<td>I do my best to assure patients'/visitors' needs are met.</td>
</tr>
<tr>
<td>35</td>
<td>0.999</td>
<td>I am mindful of patients' privacy and dignity.</td>
</tr>
<tr>
<td>36</td>
<td>0.918</td>
<td>I respond with appropriate emotions when interacting with patients/visitors.</td>
</tr>
<tr>
<td>37</td>
<td>0.922</td>
<td>I am sincere in my apologies.</td>
</tr>
<tr>
<td>38</td>
<td>0.892</td>
<td>I acknowledge and respect patients'/visitors' emotions.</td>
</tr>
<tr>
<td>39</td>
<td>0.861</td>
<td>I do my best to show compassion for patients or families.</td>
</tr>
<tr>
<td>40</td>
<td>0.875</td>
<td>I encourage patients/visitors to ask for help when necessary.</td>
</tr>
<tr>
<td>41</td>
<td>0.949</td>
<td>When I forward a patient's/visitor's concern, I follow up to ensure it has been handled.</td>
</tr>
<tr>
<td>42</td>
<td>0.970</td>
<td>When I apologize, I take care not to place blame.</td>
</tr>
<tr>
<td>43</td>
<td>0.950</td>
<td>I can clarify or explain what patients/visitors do not understand.</td>
</tr>
<tr>
<td>44</td>
<td>0.907</td>
<td>I strive to be professional in my interactions.</td>
</tr>
<tr>
<td>45</td>
<td>0.894</td>
<td>I recognize that each patient/visitor has a unique experience and circumstances.</td>
</tr>
<tr>
<td>46</td>
<td>0.972</td>
<td>I close every interaction by saying &quot;Thank you&quot; to patients/visitors.</td>
</tr>
</tbody>
</table>
References


*Journal of Services Marketing, 9*, 22-35.


