Ist der Kunde wirklich König?

Empirische Analyse der Bedingungen und Konsequenzen von Service-Qualität und Entwicklung eines computerbasierten adaptiven Tests zur ökonomischen Messung von Service-Qualität in deutschen Unternehmen.

Vorgelegt von:

Andreas Pfeiffer

Vom Promotionsausschuss des Fachbereichs 8: Psychologie der Universität Koblenz-Landau zur Verleihung des akademischen Grades Doktor der Philosophie (Dr. phil.) genehmigte Dissertation

Datum der wissenschaftlichen Aussprache:
9. Juli 2021

Vorsitzenden des Promotionsausschusses:	Prof. Dr. Ingmar Hosenfeld
Berichterstatter:	Prof. Dr. Ottmar L. Braun
	Prof. Dr. Andreas Schöler

Inhaltsverzeichnis

1 Zusammenfassung

In Kooperation mit dem TÜV SÜD und 985 Führungskräften aus deutschen Unternehmen wurde erprobt, wie Service-Qualität im Rahmen einer Onlinebefragung gemessen werden kann. Es wurde untersucht, welche Komponenten Service-Qualität umfasst, und ein Rahmenmodell entwickelt, das die Zusammenhänge zwischen Service-Qualität, Kundenzufriedenheit und dem Erfolg von Organisationen beschreibt. Die theoretische Konzeption und Operationalisierung des Konstrukts wurde mittels konfirmatorischer Faktorenanalysen überprüft und bestätigt. Das Rahmenmodell der Studie wurde als Strukturgleichungsmodell formuliert und konnte ebenfalls empirisch bestätigt werden. Die Ergebnisse und deren Auswirkung auf die Weiterentwicklung der bestehenden wissenschaftlichen Theorien zu den Konstrukten Service-Qualität, Kundenzufriedenheit und Erfolg von Organisationen wurden kritisch diskutiert.

Zur Steigerung der Ökonomie des Verfahrens wurde ein adaptiver Test zur Erfassung von Service-Qualität entwickelt. Im Rahmen der probabilistischen Testtheorie wurde geprüft, welches Item Response-Modell die empirischen Daten gut beschreiben kann. Als Grundlage für den adaptiven Test wurden die Item-Parameter modellkonform bestimmt. In einer Simulationsstudie wurde untersucht, ob die Ergebnisse der Onlinebefragung sich bedeutsam von den Ergebnissen adaptiver Tests mit unterschiedlichen Konfigurationen unterscheiden. Der Vergleich der Konfigurationen, die sich darin unterschieden, wie der Personenparameter geschätzt wurde und nach welchem Algorithmus das nächste Test-Item gewählt wurde, zeigte, welche Konfigurationen eingesetzt werden können, um eine möglichst geringe Testlänge zu erzielen, ohne dabei bedeutsame Einbußen bei der Reliabilität und Validität der Messung in Kauf zu nehmen. Unter Berücksichtigung dieser Erkenntnisse wurde der Fragebogen zur Erfassung von Service-Qualität als computerbasierter adaptiver Test umgesetzt. Diese neue Erfassungsmethode wurde in der Praxis erprobt, und abschließend wurden Nützlichkeit, Ökonomie und mögliche Nachteile, die mit dieser Art des Testens verknüpft sind, diskutiert.

Dieses Dokument, detailierte Analyseergebnisse, den adaptiven Test zur Erfassung von Service-Qualität und weitere Begleitmaterialien finden Sie unter:

- https://promotion.creaval.de

2 Einleitung

Über die Service-Qualität in Deutschland wird häufig geklagt. Schlechte Beratung, schlechter Umgang mit Beschwerden, Unfreundlichkeit im Kundenkontakt, lange Wartezeiten und schlecht qualifiziertes Personal sind nur einige Beispiele für Erfahrungen, von denen Kundinnen und Kunden berichten. Wie passen solche Erfahrungen mit der Philosophie vom ‚König Kunde' zusammen (Frey, 1997)? Wie steht es um die Service-Qualität in Deutschland wirklich? Um Antworten auf diese Fragen zu finden, wurde Service-Qualität mit Hilfe von wissenschaftlichen empirischen Methoden messbar gemacht und untersucht, von welchen Faktoren Service-Qualität abhängt und in welchem Zusammenhang sie mit Kundenzufriedenheit und dem Erfolg von Organisationen steht.

Der erste Teil dieser Arbeit umfasst eine Literaturanalyse, die die verschiedenen Komponenten von Service-Qualität unter die Lupe nimmt und zusammenfasst, welche wissenschaftlichen Erkenntnisse zu den Bedingungen und Konsequenzen von Service-Qualität vorliegen. Aufbauend auf der Darstellung des aktuellen Stands der Forschung zum Thema Service-Qualität wurde ein Rahmenmodell entwickelt, das sowohl die Komponenten von Service-Qualität umfasst, als auch die Zusammenhänge zwischen Service-Qualität, Kundenzufriedenheit und dem Erfolg von Organisationen spezifiziert. Anhand empirischer Daten aus deutschen Unternehmen wurde geprüft, ob die Operationalisierung von Service-Qualität mit dem in dieser Studie entwickelten Befragungsinstrument gelungen ist. Zudem wurde das im theoretischen Teil erarbeitete Rahmenmodell anhand der gewonnenen Stichprobe geprüft. Die Ergebnisse der Analyse werden dargestellt und kritisch beurteilt. Der erste Teil dieser Arbeit endet mit einer Diskussion, in der neben den Ergebnissen auch die Herangehensweise dieser Arbeit kritisch hinterfragt wird. Die empirischen Ergebnisse werden mit dem aktuellen Stand der wissenschaftlichen Theoriebildung verglichen und resümiert, welche Implikationen sie für die aktuelle und zukünftige Forschung im Bereich Service-Qualität haben.

Im zweiten Teil dieser Arbeit wird beschrieben, wie ein adaptiver Test zur Erfassung von Service-Qualität entwickelt wurde. Dazu wird aus der Perspektive der probabilistischen Testtheorie geprüft, wie gut die Messung von Service-Qualität mit den im ersten Teil der Arbeit entwickelten Fragebogen-Items gelingt. Eine Simulationsstudie untersucht, welche Konfiguration eines adaptiven Tests es ermöglicht, die Ökonomie bei der Erfassung des Konstrukts zu steigern, ohne dabei bedeutsame Einbußen in Bezug auf die Reliabilität und Validität der Diagnostik in Kauf zu nehmen (Moosbrugger & Kelava, 2011). Aufbauend auf den Ergebnissen dieser Simulationsstudie wurde ein webbasierter, adaptiver Test zur Erfassung von Service-Qualität entwickelt und erprobt. Im Rahmen einer Diskussion wird kritisch hinterfragt, ob die

Methode des adaptiven Testens in diesem Kontext bedeutsame Vorteile mit sich bringt und zukunftsweisend ist. Ein Ausblick, der den zweiten Teil dieser Arbeit abschließt, gibt Hinweise auf Herausforderungen und weitere lohnenswerte Fragestellungen, die in künftigen Studien beachtet werden sollten.

3 Theorieteil

Bevor der theoretische Rahmen und die Fragestellung dieser Arbeit in Kapitel 3.4 hergeleitet wird, zeigt das einführende Kapitel 3.1 auf, welche Relevanz das Thema Service-Qualität für die wissenschaftliche Auseinandersetzung innerhalb der Organisationspsychologie und für die Praxis in Organisationen hat. Im anschließenden Kapitel 3.2 wird gezeigt, wie unterschiedlich das Konstrukt Service-Qualität von verschiedenen Forschergruppen und Fachrichtungen verstanden wird. Zudem wird das Konstrukt in diesem Kapitel in das Netzwerk etablierter Konstrukte eingeordnet. In Kapitel 3.3 wird darauf eingegangen, wie sich das Verständnis des Konstruktes in verschiedenen Ansätzen der Operationalisierung niederschlägt. Es werden dabei die gängigsten Fragebögen zur Erfassung von Service-Qualität vorgestellt und es wird kritisch hinterfragt, wie das Konstrukt am besten gemessen werden kann. Nach der Darlegung, wie vielschichtig die Vorstellungen und Operationalisierungen des Konstruktes sind, wird in Kapitel 3.3.5 der aktuelle Stand der Forschung zusammengefasst. Dabei wird auf zentrale Studien, die sich mit den Bedingungen und Konsequenzen von Service-Qualität befassen, eingegangen. Aufbauend auf dieser Grundlage werden in Kapitel 3.4 eigene Überlegungen zum Konstrukt Service-Qualität und zu weiteren zentralen Variablen der Organisationspsychologie mit den bisherigen Erkenntnissen der Forschung verknüpft und ein Rahmenmodell für diese Arbeit vorgestellt.

3.1 Ist das Konstrukt Service-Qualität relevant?

Um die Wichtigkeit des Themas Service-Qualität einschätzen zu können, lohnt sich ein Blick auf die Daten des Statistischen Bundesamtes, die zeigen, dass das Bruttoinlandsprodukt in Deutschland für das Jahr 2019 zu 69 % im Dienstleistungssektor erwirtschaftet wurde. Damit dominiert der Service-Bereich die deutsche Wirtschaft und liegt weit vor dem Sektor des produzierenden Gewerbes, der als zweitwichtigster Wirtschaftsbereich mit 24 % zum Bruttoinlandsprodukt beitrug (Statistisches Bundesamt, 2020).

Weitet man den Blick über die Grenzen Deutschlands aus, so stellt man fest, dass der Anteil des Dienstleistungssektors an der Gesamtwirtschaftsleistung von Staaten europaweit und in den meisten Industrienationen in dieser Größenordnung liegt (Central Intelligence Agency, 2020). Damit wird deutlichen, dass Dienstleistungen heutzutage in allen bedeutenden Wirtschaftsnationen den größten Teil der erbrachten Wirtschaftsleistungen ausmachen. Die Tatsache, dass Dienstleistungen mittlerweile den zentralen Wirtschaftszweig der meisten Staaten darstellen,

wird durch das Phänomen der Tertiarisierung beschrieben, das den in den 1970er-Jahren beginnenden strukturellen Wandel hin zu Dienstleistungsgesellschaften bezeichnet (Häussermann, 1995; Klodt, Maurer & Schimmelpfennig, 1997; Sozialakademie Dortmund, 1999).

Betrachtet man die Zuordnung von Unternehmen zu den gängigen Wirtschaftsbereichen Landwirtschaft, Industrie und Dienstleistungen, wird deutlich, dass diese Kategorisierung nicht für jede Organisation problemlos eindeutig möglich ist. So werden beispielsweise in der Automobilindustrie als Produkt primär Autos gebaut, dieses angebotene Produkt ist jedoch vollständig eingebettet in eine Reihe von Service-Leistungen wie zum Beispiel Reparatur, Wartung und Verkauf von Autos. Es wird deutlich, dass auch in der klassischen Industrie Service-Leistungen eine immer größere Rolle spielen und einen erheblichen Teil der Unternehmensleistung darstellen.

Üblicherweise wird bei Service-Qualität an die Schnittstelle von Organisationen mit ihrer Kundschaft gedacht. Ein Blick auf die Strukturen innerhalb von Organisationen zeigt, dass auch organisationsintern Kunden-Dienstleister-Beziehungen beobachtet werden können. So unterstützt z. B. eine IT-Abteilung andere Unternehmenszweige mit internen Dienstleistungen. Nimmt man diese Perspektive ein, wird deutlich, wie wichtig eine qualitativ hochwertige Erbringung solcher interner Service-Leistungen ist, um die gesetzten Ziele zu erreichen.

Wie bereits in der Einleitung beschrieben, wird der Begriff Service-Qualität im Management heutiger Organisationen großgeschrieben. Ruft man beim Kundendienst großer Telefonie- und Internetanbieter an, wird man meist, bereits bevor man mit dem ersten Menschen spricht, von einer Computerstimme gefragt, ob das nachfolgende Gespräch zur Verbesserung der Service-Qualität aufgezeichnet werden darf. Nicht nur solche großen Serviceanbieter messen dem Thema eine hohe Bedeutung bei, auch in der Verwaltung von Kommunen ist das Thema Service-Qualität immer häufiger Inhalt von Mitarbeiterbefragungen und Dienstbesprechungen.

Vor über 25 Jahren erkannte Schneider (1990b, S. 389): „Many service organizations are beginning to view service quality or service excellence as a strategic imperative or, at minimum, a strategic opportunity“ und prognostizierte damit eine Entwicklung, deren Ergebnisse wir heute beobachten können. Service-Qualität wurde auch als Chance gesehen, um Marktdominanz zu erreichen. „Many companies have discovered that attention to the softer side of business is a way of achieving competitive dominance in their markets“ (Schneider, Holcombe & White, 1997, S. 35). Um in immer dichter besetzten Märkten an der Spitze der erfolgreichen Organisationen zu bleiben, ist exzellenter Service heutzutage eine Grundvoraussetzung.

Bei allen Vorteilen, die mit hoher Service-Qualität verbunden sind, lohnt auch ein Blick auf die Kosten, die mit schlechtem Service verknüpft sind (Heskett, Sasser & Hart, 1991). Personen, die die Erfahrung gemacht haben, dass der Service einer Organisation nicht ihren Erwartungen entspricht, geben diese Erfahrung in ihrem Freundes- und Bekanntenkreis, aber auch zunehmend in Bewertungsportalen im Internet weiter. Wie groß der Schaden für eine Organisation

ausfällt, wenn ihre Kundschaft offen kommuniziert, dass sie von der Service-Qualität enttäuscht ist, lässt sich schwer in Zahlen fassen. In vielen Bereichen, wie z. B. bei Ärzten, Rechtsanwälten und Restaurants, spielt bei der Wahl des Anbieters die persönliche Weiterempfehlungen eine große Rolle. In diesen Bereichen ist es offensichtlich, dass ein negatives Service-Image in der Öffentlichkeit verheerende Auswirkungen haben kann.

Wie in Kapitel 3.3.5 ausführlich beschrieben wird, gibt es zahlreiche Studien, die den Zusammenhang zwischen Service-Qualität und Kundenzufriedenheit belegen (Brady & Robertson, 2001; Brady & Cronin Jr, 2001b; Cronin Jr, Brady & Hult, 2000; Schneider et al., 1997). Auch der Zusammenhang zwischen Kundenzufriedenheit und Wiederkaufbereitschaft bzw. der Bereitschaft, ein Unternehmen weiter zu empfehlen, ist empirisch ausreichend belegt (Braun & Müssigmann, 2009b; Rust & Zahorik, 1993). Glaubt man an eine kausale Verknüpfung dieser zentralen Variablen in einer Organisation, so sollte dem Thema Service-Qualität besondere Aufmerksamkeit gewidmet werden.

Weil Service-Qualität im Vergleich zu ähnlichen Konstrukten wie zum Beispiel Produktqualität ein relativ junges Feld ist und weil herausragend guter Service nicht so einfach sichergestellt werden kann, ist es wichtig, genauer zu verstehen, was Service-Qualität ausmacht und wie das Management von Organisationen diese sicherstellen kann. Wenn Service-Qualität ein Eckstein der Unternehmensstrategie ist, wird ein gutes Instrument zur Messung von Service-Qualität benötigt, denn ohne eine reliable, valide und ökonomische Diagnostik ist wirkungsvolles Service-Management und die damit verbundene Auswahl bzw. Entwicklung angemessener Interventionen sowie die Kontrolle von deren Wirksamkeit nicht möglich.

3.2 Service-Qualität und benachbarte Konstrukte

Das Konstrukt Service-Qualität stellt für die psychologische Forschung in vielfacher Sicht eine Herausforderung dar. Dies lässt sich unter anderem daran erkennen, dass es in den Publikationen des Fachs bislang keine einheitliche Definition des Konstruktes gibt. Bevor in Kapitel 3.3 einige theoretische Konzeptionen und Definitionen für das Konstrukt vorgestellt und diskutiert werden, soll in diesem Kapitel aufgezeigt werden, wie Service-Qualität in das Geflecht anderer Konstrukte eingebettet werden kann.

Zunächst wird dabei auf Produktqualität eingegangen, da dieses benachbarte Konstrukt in den Ingenieurwissenschaften, aber auch in zahlreichen Managementschulen bereits ausführliche Berücksichtigung gefunden hat (Hinrichs, 2005). Man könnte zunächst annehmen, dass Ansätze aus dem Bereich Produktqualität leicht auf das Thema Service-Qualität übertragbar sind, weshalb die wichtigsten Ansätze aus diesem Bereich beschrieben werden. In Kapitel 3.3 wird anschließend beschrieben, warum sich diese Ansätze nur eingeschränkt auf die Qualität von Dienstleistungen übertragen lassen. Als weitere benachbarte Konstrukte, die je nach

deren Verständnis große Schnittmengen mit dem Konstrukt Service-Qualität haben, werden in diesem Kapitel Service-Klima, Service-Orientierung, Kundenorientierung und Kundenzufriedenheit vorgestellt.

3.2.1 Produktqualität

Bei der Produktion von Waren wird unter dem Begriff Qualität und der Qualitätssicherung verstanden, wie exakt ein Produkt den vorgegeben Spezifikationen entspricht (Golder, Mitra & Moorman, 2012). Neben diesem Qualitätsverständnis, das auch als Produktqualität bezeichnet wird und sich primär mit dem fertigen Produkt befasst, kann auch die Prozessqualität betrachtet werden, bei der es um die Zuverlässigkeit des Herstellungsprozesses geht.

Sowohl Produkt- als auch Prozessqualität sind sehr etablierte Themenfelder, denen bereits von verschiedenen wissenschaftlichen Disziplinen Aufmerksamkeit geschenkt wurde (Braun & Koch, 2002; Winz, 2016). Im Folgenden wird kurz auf die zentralen Konzepte der Auseinandersetzung mit diesen Themen eingegangen und abschließend gezeigt, wo die Nahtstellen zwischen Produktqualität und Service-Qualität verlaufen.

Zur Steigerung der Produktqualität wurden in vielen Unternehmen Qualitätszirkel installiert. Da die Umsetzung von Qualitätszirkeln in verschiedenen Organisationen sehr unterschiedlich ausfällt, lässt sich keine einheitliche Definition des Begriffes geben (Bungard, 1991). In den meisten Umsetzungen bestehen Qualitätszirkel aus regelmäßigen ein- bis zweistündigen Gesprächsrunden von fünf bis zehn Organisationsmitgliedern, die teilweise aus verschiedenen, meistens jedoch aus den unteren Hierarchieebenen einer Produktionsstätte stammen. Die Treffen finden in der Regel während der Arbeitszeit statt, werden von Angestellten oder Führungskräften moderiert und dienen primär der Steigerung der Produktqualität. Im Rahmen der Gruppengespräche wird großer Freiraum eingeräumt, so dass alle Beteiligte ihr Wissen einbringen und selbstständig Lösungsvorschläge erarbeiten können, die häufig im Rahmen des betrieblichen Vorschlagswesens belohnt werden. Durch den angebotenen Handlungsspielraum soll die Motivation und die Identifikation mit dem Unternehmen gefördert werden (Bungard, 1992). Ähnliche Konzepte existieren schon seit längerer Zeit in der Automobilindustrie und werden als „Werkstattzirkel“ oder „Lernstatt“ bezeichnet. Ackermann (1992) untersuchte die Rahmenbedingungen, unter denen Qualitätszirkel besonders effizient sind, und konnte feststellen, dass die Auswahl und intensive Ausbildung geeigneter Moderatoren, die Unterstützung durch Management und direkte Vorgesetzte, die Langfristigkeit und Regelmäßigkeit der Gruppensitzungen sowie die Freiwilligkeit der Teilnahme und das Mitwirken der Gruppenmitglieder entscheidende Erfolgsfaktoren sind.

Ein weiterer Ansatz, der darauf abzielt, kürzere Durchlauf- und Entwicklungszeiten, eine gesteigerte Produktivität und eine Steigerung der Produktqualität zu erreichen, ist das Lean

Management (Müller & Rupper, 1994; Wildemann, 1999). Der Ursprung dieses Managementkonzepts ist das International Motor Vehicle Program am Massachusetts Institute of Technology von Womack, Jones und Roos (1994), in dem das Schlagwort lean production, das als „schlanke Produktion" übersetzt werden kann, geprägt wurde. Womack et al. (1994) verglichen die Produktion und Entwicklung von Autos und stellten fest, dass japanische Autohersteller, die nach dem Zweiten Weltkrieg mit der zerstörten japanischen Wirtschaft und knappen Ressourcen auskommen mussten, Produktionssysteme entwickelten, die qualitativ hochwertige Produkte hervorbrachten und zudem geringere Kosten verursachten. Hauptorganisationsmerkmale dieser schlanken Produktion ist die Übertragung von Aufgaben und Verantwortung an die Mitarbeiter, die Installation von Systemen zur Fehlerentdeckung, Gruppenarbeit sowie die Optimierung der Zulieferkette (Stürzl, 1992). Lean Management bzw. Lean Production zielen auf hohe Produktqualität ab, konzentrieren sich bei der Umsetzung jedoch sehr auf den Produktionsprozess und unterstreichen damit die Wichtigkeit der Prozessqualität.

Bei der Umsetzung eines Qualitätsmanagementsystems bieten die der von der Deutschen Gesellschaft für Qualität vorgelegten DIN EN ISO 9000-Normen eine gute Orientierung im Hinblick auf technische Abläufe bzw. technische Systeme (DIN EN ISO 9000:2015-11). Zahlreiche Unternehmen haben sich an diesen Normen orientiert und durchliefen einen Zertifizierungsaudit durch staatlich zugelassene, unabhängigen Stellen, um anhand eines Zertifikates nachweisen zu können, dass sie ein wirksames System zur Qualitätssicherung etabliert haben (Brauer & Kühme, 1996).

Mit dem in den 90er-Jahren immer stärker werdenden Qualitätswettbewerb wurde das Total Quality Management (TQM) als unternehmerisches Erfolgskonzept entwickelt und immer weiter verbreitet. Der Ansatz des TQM ist ein umfassender Denk- und Handlungsansatz, der Qualitätsbewusstsein und Qualitätssicherung in allen Phasen der Wertschöpfungskette und bei allen Führungskräften und Angestellten umfasst (Töpfer & Mehdorn, 1995). TQM bezieht neben der Qualität 1. Grades, der technischen Produktqualität, auch die sogenannte Qualität 2. Grades mit ein, die alle Kontaktphasen mit dem Kunden umfasst und Kundenzufriedenheit als Maßstab hat. Damit umfasst TQM auch Elemente der Service-Qualität.

In den Normen der ISO 9000-Gruppe sowie im TQM werden Kundinnen und Kunden mit deren Anforderungen und Erwartungen berücksichtigt. Bovermann (2013) zeigt, dass TQM einen guten Rahmen bietet, um den Weg zu höherer Dienstleistungs- bzw. Service-Qualität zu bahnen. Sie fordert, dass umfassendes TQM die Planung von Serviceleistungen umfassen muss und liefert mit ihrem Regelkreis der Dienstleistungsqualität ein Managementkonzept, um Dienstleistungsqualität ausgehend von den Erwartungen der Kunden zu steuern.

Powell (1995) untersuchte, ob Unternehmen die TQM nutzen bzw. schon länger einsetzen, um ihre Konkurrenten hinsichtlich der Unternehmensleistung übertreffen. Zudem untersuchte er den korrelativen Zusammenhang zwischen TQM und einer Reihe von Parametern

aus der Arbeitswelt wie zum Beispiel: Führung, Kommunikation, Beziehungen zu Kunden und Lieferanten, flexibler Fertigung und Prozessverbesserungen. Seine Ergebnisse zeigen, dass TQM zum wirtschaftlichen Nutzen beitragen kann, dies jedoch nicht in allen Organisationen gleichermaßen tut. Als entscheidende Faktoren identifiziert er die engagierte und verbindliche Umsetzung des Konzepts sowie die Offenheit der Organisation für das Konzept.

TQM umfasst das Thema Kontinuierlicher Verbesserungsprozess (KVP), das aus der japanischen Lebens- und Arbeitsphilosophie Kaizen abgeleitet wurde (Steinbeck, 1995). Wesentliche Elemente des KVP sind die Entwicklung von Instrumentarien für dauerhafte und andauernde Verbesserungen in allen Bereichen. Dieser Ansatz umfasst die Personalführung mit dem Ziel der Optimierung von Verfahren und Produkten sowie die Idee der Verhaltenssteuerung durch Ziele, Anreize und Ergebnisrückkopplung.

Wie aus der bisherigen Darstellung erkennbar wird, hängen die Schlagworte aus dem Bereich Produktqualität: Lean Production, Kaizen, TQM und KVP eng zusammen und können in der praktischen Umsetzung schwer voneinander getrennt werden. Auch Hegner (1994) kommt zu diesem Schluss und fasst zusammen, dass die Kernbestandteile dieser Managementkonzepte altbekannte Merkmale der Organisationsentwicklung sind, die die Innovation von Produkten und Produktionstechnologien durch ständige kleine Verbesserungen und unter der Beteiligung aller Betroffenen ermöglichen.

Wie in diesem Abschnitt deutlich wurde, ist die Qualität von Produkten Inhalt verschiedener Theorien und Managementstrategien. Da sich Service-Leistungen in vielerlei Hinsicht entscheidend von Produkten unterscheiden, kommen Parasuraman, Berry und Zeithaml (1991b, S. 253) zu der Schlussfolgerung: „Knowledge about goods quality is insufficient to understand service quality." Welche Unterschiede zu der eingeschränkten Übertragbarkeit der Erkenntnisse aus dem Bereich Produktqualität auf das Thema Service-Qualität führen, wird in Kapitel 3.3 detailliert beschrieben.

3.2.2 Service-Klima

Das Konstrukt Service-Klima hat sich aus dem Umfeld des Konstruktes Organisationsklima bzw. Organisationskultur entwickelt, was sich sehr deutlich an der Ähnlichkeit der Definition der beiden Konstrukte erkennen lässt. Schneider (1990a, S. 384) definiert Organisationsklima als „incumbents' perceptions of the events, practices, and procedures and the kinds of behaviors that get rewarded, supported, and expected in a setting". Service-Klima wird von Schneider, White und Paul (1998, S. 151) als „employee perceptions of the practices, procedures, and behaviors that get rewarded, supported, and expected with regard to customer service and customer service quality" definiert. Ein Vergleich der Definitionen verdeutlicht, dass Service-Klima die Anwendung des Konstruktes Organisationsklima auf den Inhaltsbereich Service darstellt.

Das Konstrukt Organisationsklima lässt sich ebenso gut auf andere Bereiche wie zum Beispiel Sicherheit oder Innovation anwenden (Ashkanasy, Wilderom & Peterson, 2000; Schneider, 1990b). Gemäß der Definition von Service-Klima zeichnen sich Organisationen mit einem starken Service-Klima dadurch aus, dass ihre Mitglieder das Gefühl haben, für exzellenten Service belohnt zu werden, und wahrnehmen, dass guter Service für das Management von hoher Bedeutung ist. Eine weitere ähnliche Definition des Konstruktes findet sich bei Auh, Menguc, Fisher und Haddad (2011, S. 428), die Service-Klima als „employee's perception of the extent to which the organization emphasizes excellence in customer service by providing rewards, recognition, and organizational resources such as training, tools, skills, and knowledge" beschreiben.

In der Literatur lassen sich noch einige weitere Definitionen des Konstruktes finden, die alle sehr ähnlich sind und gemeinsam haben, dass Service-Klima als gemeinsames Verständnis konzipiert wird, das durch die Interpretation und Interaktion mit sozialen Hinweisreizen aus dem Arbeitsumfeld entwickelt wird. Obwohl die Definitionen des Konstrukts sich sehr stark gleichen, fallen die Operationalisierungen des Konstruktes in Form von Fragebogenitems in verschiedenen Studien sehr unterschiedlich aus.

In einer Studie von Schneider und Bowen (1985), in der die Autoren auf faktorenanalytischen Ergebnissen vorangegangener Studien aufbauen, wird davon ausgegangen, dass das Konstrukt, wird es bei Organisationsmitgliedern erfasst, vier Subdimensionen aufweist, eingeschätzt durch die Kundschaft jedoch fünf Subdimensionen. In dieser frühen Studie wurden neben der Anzahl der Subdimensionen auch die Bezeichnungen der Subdimensionen noch einmal überarbeitet. Schneider et al. (1998, S. 153) konzipieren das Konstrukt Service-Klima anhand der vier Dimensionen „Global Service Climate", „Customer Orientation", „Managerial Practices" und „Customer Feedback", deren hohe Reliabilität sich empirisch mehrfach bestätigen ließ. In der Praxis eingesetzte Fragebögen zur Erfassung von Service-Klima unterscheiden sich bei der Formulierung der Items. Häufig wurde versucht, diese möglichst generisch zu halten, so dass die entwickelte Skala in möglichst vielen verschiedenen Organisationen und Branchen eingesetzt werden kann, in anderen Anwendungsfällen sollten die Items spezifische Aspekte einer Organisation abbilden, um höhere Akzeptanz zu erreichen. Aktuelle Publikationen zum Thema Service-Klima gehen von einem homogenen Konstrukt aus, das keine Subdimensionen aufweist (Bowen & Schneider, 2014).

Welchen Einfluss die Persönlichkeitsdimensionen aus dem Fünf-Faktoren-Modell der Persönlichkeitspsychologie auf die Wahrnehmung des Service-Klimas haben, untersuchten Auh et al. (2011). Sie analysierten ihre Daten mit Mehrebenenmodellen und stellten fest, dass Mitarbeiterinnen und Mitarbeiter, die hohe Werte auf den Persönlichkeitsdimensionen Gewissenhaftigkeit, Offenheit für Erfahrungen und Verträglichkeit haben, das Service-Klima positiver wahrnehmen. Ein durch das Personal positiv eingeschätztes Service-Klima steht gemäß ihren

Ergebnissen in positivem Zusammenhang mit der Zufriedenheit der Kundschaft und deren Entscheidung, das Geschäft besucht zu haben.

Mittlerweile liegen zahlreiche Publikationen zum Konstrukt Service-Klima vor, die sich mit den Bedingungen, den Konsequenzen oder moderierenden Effekten von Service-Klima befassen (Brady & Cronin Jr, 2001a; Dean, 2004; Jong, Ruyter & Lemmink, 2004; Ehrhart, Witt, Schneider & Perry, 2011; Hartline & Ferrell, 1996; Salanova, Agut & Peiro, 2005; Schneider & Bowen, 1985; Schneider, 1990a; Schneider et al., 1998; Schneider, Bowen, Ehrhart & Holcombe, 2000; Schneider, Salvaggio & Subirats, 2002; Schneider, Macey, Lee & Young, 2009; Schneider, Ehrhart & Macey, 2013; Yagil, 2001).

Die Befunde dieser Einzelstudien haben Hong, Liao, Hu und Jiang (2013) zusammengefasst und in ein theoretisches Modell integriert, das in Abbildung 1 dargestellt wird.

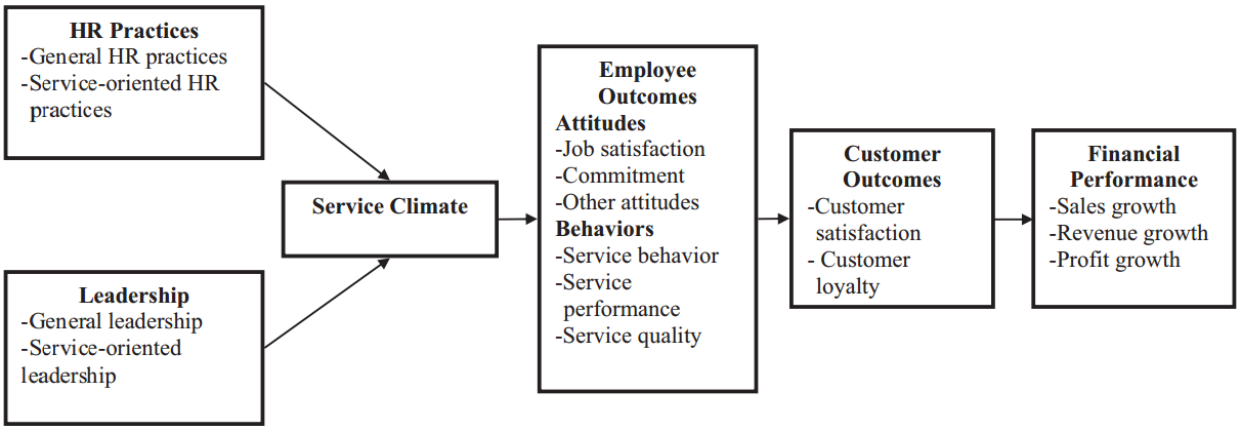


Abbildung 1 Modell der Bedingungen und Konsequenzen von Service-Klima aus (Hong et al., 2013)

In ihrer Arbeit ordnen sie Service-Klima als zentrales Bindeglied in den weiteren Rahmen der Service Profit Chain ein (Heskett, Sasser & Schlesinger, 1997). Als wesentliche Bedingungen für Service-Klima benennen sie das Personalwesen und die Führung in der Organisation. Service-Klima führt zu verschiedenen Einstellungen und Verhaltensweisen des Personals. Für diese Arbeit ist von besonderem Interesse, dass Hong et al. (2013) Service-Qualität als eine Konsequenz von Service-Klima konzipieren. Auch die weiteren Annahmen ihres Modells, dass diese Einstellungen und Verhaltensweisen der Angestellten sich auf die Kundschaft auswirken und unter anderem zu Kundenzufriedenheit und Kundenbindung führen, sowie die Annahme, dass dadurch der wirtschaftliche Erfolg einer Organisation beeinflusst wird, stellen zentrale Annahmen in dieser Arbeit dar (s. Kapitel 3.4). Um ihr theoretisch hergeleitetes Modell empirisch

zu prüfen, führten Hong et al. (2013) eine Metaanalyse durch, die ihr Modell auf korrelativer Ebene und modelliert als Pfadmodell untermauert.

Einen weiteren umfassenden Überblick zum Konstrukt Service-Klima, der in Abbildung 2 wiedergegeben wird, publizierten Bowen und Schneider (2014).

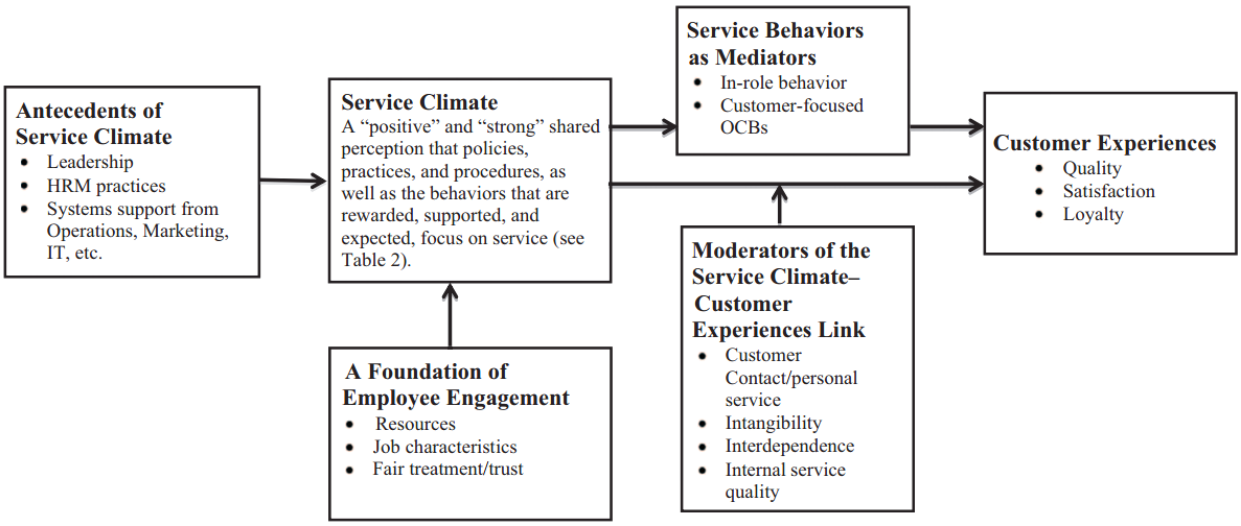


Abbildung 2 Rahmenmodell des Konstruktes Service-Klima aus (Bowen & Schneider, 2014)

Ihr Rahmenmodell enthält die gleichen Konstrukte als Bedingung für Service-Klima wie das Modell von Hong et al. (2013) und benennt darüber hinaus das Engagement der Arbeitskräfte als notwendige, aber nicht hinreichende Bedingung für positives Service-Klima. Bowen und Schneider (2014) gehen davon aus, dass Service-Klima einen direkten und einen indirekten, über Service-Verhaltensweisen vermittelten Effekt auf das Erleben der Kundschaft hat. Für den direkten Effekt von Service-Klima auf Kundenzufriedenheit, Kundenbindung und die Qualität der Kundenerfahrung benennen sie Moderatoren wie zum Beispiel die interne Service-Qualität. Das formulierte Modell basiert auf zahlreichen Studien, in denen Zusammenhänge untersucht wurden, die auch im formulierten Gesamtmodell enthalten sind. Eine empirische Studie, in der alle Elemente des Modells erfasst wurden, so dass das gesamte Modell anhand eines Datensatzes geprüft werden kann, liegt bislang nicht vor.

Das Konstrukt Service-Klima wird meistens auf der Individualebene erfasst, einzelne Teammitglieder werden befragt, wie sie das Service-Klima in ihrem Arbeitsumfeld wahrnehmen (Auh et al., 2011; Liao & Chuang, 2007; Liao & Subramony, 2008). Diese Herangehensweise ist sinnvoll, wenn untersucht werden soll, wie zum Beispiel Persönlichkeitseigenschaften der Teammitglieder und deren Einschätzung des Service-Klimas zusammenhängen. In zahlreichen Studien wurde der Zusammenhang zwischen Service-Klima und Variablen wie zum Beispiel

Servicestrategie, Führungsverhalten oder Indikatoren der wirtschaftlichen Leistungsfähigkeit wie Umsatz oder Gewinn untersucht (Chen, Zhu & Zhou, 2015; Hui, Chiu, Yu, Cheng & Tse, 2007; Salvaggio et al., 2007; Schneider & Bowen, 1985). Da diese Variablen nicht von Teammitglied zu Teammitglied variieren, sondern für Teammitglieder einer Organisationseinheit konstant sind, kann man sie im Rahmen von Mehrebenenanalysen als Variablen auf einer höheren Ebene betrachten. Untersucht man beispielsweise wie Hui et al. (2007) in sechs Organisationen, wie sich das Führungsverhalten von 55 Führungskräften auf die Einschätzung des Service-Klimas durch 511 Angestellte auswirkt, sollte diese hierarchische Struktur der Daten berücksichtigt werden. Um das Führungsverhalten in Zusammenhang mit dem Service-Klima zu untersuchen, werden die Einschätzungen der einzelnen Beschäftigten meistens gemittelt und der resultierende Wert als Indikator für das Service-Klima der entsprechenden Organisationseinheit genutzt. Wie Dietz, Pugh und Wiley (2004) zeigen konnten, ist diese gängige Praxis möglicherweise problematisch, weil Angestellte durchaus unterschiedliche Einschätzungen zum Service-Klima in der Niederlassung und zum Service-Klima in der Gesamtorganisation abgeben können. Die Wichtigkeit der Mehrebenenstruktur und der daraus resultierenden geschachtelte Datenstruktur unterstreichen auch Subramony und Pugh (2015) in ihrem integrativen Modell, das in Abbildung 3 dargestellt wird.

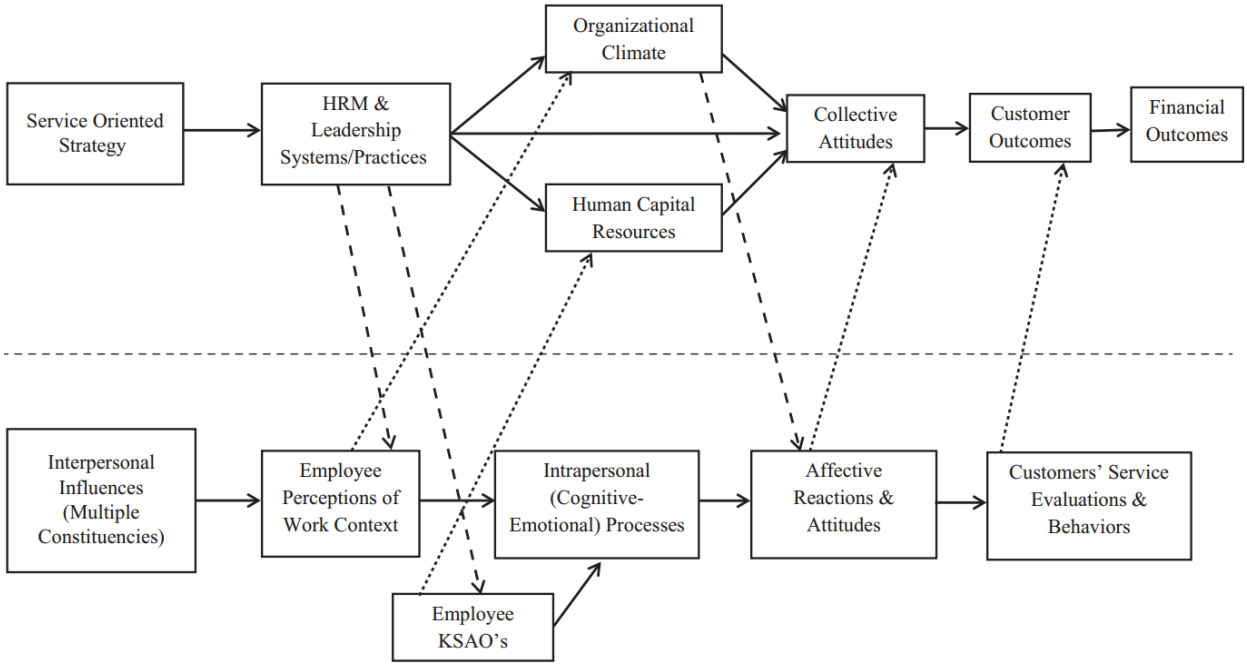


Abbildung 3 Integratives Modell für Service-Management aus (Subramony & Pugh, 2015)

In der Mitte des Modells von Subramony und Pugh (2015) markiert eine gestrichelte, horizontale Linie die Grenze zwischen der Ebene des Individuums und der Ebene von höheren Organisationseinheiten. Service-Klima als Teilaspekt von Organisationsklima wird in ihrem Modell als Konstrukt auf der Ebene von Organisationseinheiten gesehen und von Variablen auf dieser Ebene wie zum Beispiel Personalwirtschaft und Führung beeinflusst. Einen weiteren Einfluss auf das Organisationsklima haben die Wahrnehmungen des Arbeitsumfeldes durch die Beschäftigten, die in diesem Modell eine Variable darstellen, die sich unterhalb der Trennlinie befindet und damit auf der Ebene der einzelnen Person angesiedelt ist.

Wie in den berichteten Studien und Modellen deutlich wird, steht das Konstrukt Service-Klima in engem Bezug zum Thema Service-Qualität. Die Literatur zum Thema Service-Klima weist anhand praktischer Beispiele auf Probleme hin, die sich im Zusammenhang mit hierarchisch geschachtelten Datensätzen ergeben und die auch bei Studien zum Thema Service-Qualität berücksichtig werden sollten. Zudem wurde die zentrale Idee dieser Arbeit, Service-Qualität mit Kundenzufriedenheit und dem wirtschaftlichen Erfolg von Organisationen in Verbindung zu bringen, in der Literatur zum Thema Service-Klima berücksichtigt; es liegen vielversprechende Befunde vor, dass diese Verknüpfung theoretisch begründbar und empirisch nachweisbar ist.

3.2.3 Service-Orientierung

Service-Orientierung ist ein weiteres Konstrukt, das in direktem Zusammenhang mit Service-Qualität steht. Mit der Zielsetzung ein Persönlichkeitspsychologisches Maß für praktische Selektionsaufgaben im Bereich der Eignungsdiagnostik und Mitarbeiterselektion bzw. -platzierung zu entwickeln, schlugen Hogan, Hogan und Busch (1984, S. 167) das Konstrukt „Service Orientation“ als „the disposition to be helpful, thoughtful, considerate, and cooperative“ vor. Sie beschreiben das Konstrukt als „a set of attitudes and behaviors that affects the quality of the interaction between ... the staff of any organization and its customers“ (Hogan et al., 1984, S. 167). Bettencourt, Gwinner und Meuter (2001, S. 31) beschreiben Service-Orientierung als „an individual's predisposition to provide superior service through responsiveness, courtesy, and genuine desire to satisfy customer needs“. Service-Orientierung ist damit als Persönlichkeitsmerkmal konzipiert, das die Qualität des Verhaltens von Organisationsmitgliedern, die in Kontakt mit Kunden stehen, beschreibt.

Die erste Operationalisierung des Konstruktes, eine Skala bestehend aus 92 Items, die den „Service Orientation Index (SOI)“ bilden, stammt von Hogan et al. (1984). Dienhart, Gregoire und Downey (1990) nutzten neun Items, um Service-Orientierung zu messen. Mittels einer exploratorischen Hauptachsen-Faktorenanalyse extrahierten sie die drei Faktoren „Organization Support“, „Customer Focus“, und „Service Under Pressure“, die jeweils einen Eigenwert über 1 erreichten (Dienhart et al., 1990, S. 424). Die Reliabilität dieser Subskalen lag zwischen $\alpha = .31$

und α = .65 und fiel damit sehr niedrig aus, was teilweise durch die geringe Anzahl an Items pro Dimension erklärt werden konnte. Groves, Gregoire und Downey (1995) entwickelten, ausgehend von 50 Aussagen zum Thema Service-Orientierung ein Instrument, um Service-Orientierung beim Personal im Restaurant zu erfassen. Sie konnten die bereits von Dienhart et al. (1990) benannten Subskalen mittels exploratorischer Faktorenanalyse replizieren. Bettencourt et al. (2001) nutzten konfirmatorische Faktorenanalysen, Maße der Item-Trennschärfe und Reliabilität und entwickelten basierend auf den von Hogan et al. (1984) beschriebenen Dimensionen eine Skala zur Erfassung von Service-Orientierung, die aus fünf Items besteht. Um Organisationen einen Überblick über Service-Leistungen beeinflussende Elemente zu geben, entwickelten Dale und Wooler (1991) ein Modell, das als eine Eigenschaft von Organisationsmitgliedern Service-Orientierung enthält. In ihrem Ansatz setzt sich Service-Orientierung aus Kontaktfreudigkeit, Sympathie, Neugier wie Dinge funktionieren, dem angemessenen Befolgen von Regeln und guter Anpassungsfähigkeit zusammen. Die Autoren berichten keine empirischen Ergebnisse, die diese angenommene Struktur belegen, und weisen darauf hin, dass ihr Rahmenmodell, insbesondere im Hinblick auf den Detaillierungsgrad, auf den konkreten Anwendungsfall angepasst werden sollte.

Im Rahmen der Konstruktvalidierung wurden von Hogan et al. (1984), neben anderen persönlichkeitspsychologischen Skalen, Fremdeinschätzungen durch andere Teammitglieder und Arbeitsleistungsindikatoren erfasst und gezeigt, dass Service-Orientierung erwartungstreue Korrelationsmuster mit anderen bekannten und validierten Maßen der Persönlichkeit und beruflichen Präferenzen aufweist. Service-Orientierung wird als nicht technische Leistung von Arbeitskräften betrachtet und ein korrelativer Zusammenhang mit deren Gesamtarbeitsleistung konnte empirisch nachgewiesen werden. Ones, Viswesvaran und Dilchert (2005) zeigten, dass die Facetten Verträglichkeit, emotionale Stabilität und Gewissenhaftigkeit des Fünf-Faktoren-Modells der Persönlichkeitspsychologie die Grundlage von Service-Orientierung sind. Sie untersuchten die Kriteriumsvalidität von Service-Orientierung anhand einer Metaanalyse und konnten diese für das Kriterium Gesamtarbeitsleistung zu bestätigen (Roberts & Hogan, 2001). Service-Orientierung ist ein Persönlichkeitskonstrukt, das auf einem geringeren Abstraktionsniveau als die Facetten des Fünf-Faktoren-Modell der Persönlichkeitspsychologie angesiedelt ist, weshalb theoretisch begründet werden kann, dass dieses Konstrukt besser geeignet ist, tatsächliches Verhalten in Service Situationen vorherzusagen, als das allgemeinere Fünf-Faktoren-Modell (Bettencourt et al., 2001; Ones et al., 2005). Cran (1994) argumentiert, dass Service-Orientierung zentrales Element bei der Selektion von Personal im Service-Bereich sein sollte, denn wie andere Trait-Konstrukte in der Psychologie erweist sich auch diese Persönlichkeitseigenschaft als zeitlich und transsituativ stabil und lässt sich durch Mitarbeiterschulungen nur sehr schwer verändern.

Um herauszufinden, wie Service-Orientierung die von Kunden wahrgenommene Service-Qualität beeinflusst, wurde von Webber, Payne und Taylor (2012) untersucht, ob kognitives und emotionales Vertrauen eine vermittelnde Rolle spielt. Um ihr Mediationsmodell zu prüfen,

nutzen sie ein Strukturgleichungsmodell und konnten empirisch zeigen, dass kognitives Vertrauen den Zusammenhang zwischen der Service-Orientierung der Organisationsmitglieder und der durch Kundschaft eingeschätzten Service-Qualität wesentlich vermittelt.

Smith, Rasmussen, Mills, Wefald und Downey (2012) setzten das Konstrukt Service-Orientierung ein, um herauszufinden, ob der Zusammenhang zwischen Stress und der Arbeitsleistung bei Servicekräften in Restaurants durch Service-Orientierung mediiert wird. Sie konnten ihre Hypothesen, dass der Zusammenhang zwischen Stress und Arbeitsleistung für Personen mit hohe Service-Orientierung geringer ausfällt, empirisch nicht bestätigen.

Service-Orientierung als Persönlichkeitseigenschaft von Organisationsmitgliedern im direkten Kundenkontakt ist ein etabliertes Konstrukt, dessen Schnittstelle zur Service-Qualität offensichtlich ist und bereits Gegenstand einiger Studien war. Vor dem Hintergrund der Prognose von Verhalten in konkreten Service-Situationen erscheint es sinnvoll, die in der Personalselektion etablierten Persönlichkeitskonstrukte um dieses Konstrukt, das auf einem spezifischeren Abstraktionsniveau angesiedelt ist, zu erweitern. Aus der Perspektive des Service-Managements steckt hinter der Forschung zu diesem Konstrukt der klare Appell, bei der Selektion und Platzierung von Personal auf diese Persönlichkeitseigenschaft zu achten. Akzeptiert man die Auffassung, dass dieses Konstrukt zeitlich und transsituativ stabil ist, sollte gründlich überlegt werden, ob Investitionen in die Schulung von Personen, die im direktem Kundenkontakt stehen, zu verhaltenswirksamen Veränderungen führen können.

3.2.4 Kundenorientierung

Ein weiteres Konstrukt, das in enger Verwandtschaft mit dem Konstrukt Service-Qualität steht, ist das Konstrukt Kundenorientierung. In der wissenschaftlichen Literatur zu Kundenorientierung finden sich zwei wesentliche Auffassungen des Konstrukts. Zum einen wird Kundenorientierung als eine Reihe von Verhaltensweisen konzeptualisiert, die dazu dienen, die Bedürfnisse der Kundschaft zu befriedigen (Homburg, Müller & Klarmann, 2011; Korschun, Bhattacharya & Swain, 2014; Michaels & Day, 1985; Rozell, Pettijohn & Parker, 2004; Saxe & Weitz, 1982; Tadepalli, 1995). Der zweite Ansatz versteht Kundenorientierung als Oberflächen-Trait; das bedeutet, dass Kundenorientierung als Persönlichkeitseigenschaft auf einem höheren Abstraktionsniveau und damit als über die Zeit stabile Prädisposition, Kundinnen und Kunden zu unterstützen und zu bedienen, konzipiert wird (Brown, Mowen, Donavan & Licata, 2002; Donavan, Brown & Mowen, 2004; Harris, Mowen & Brown, 2005; Rod & Ashill, 2010). Im Folgenden werden diese beiden Forschungsperspektiven und die daraus entstandenen Verfahren, um Kundenorientierung zu messen, dargestellt und darauf eingegangen, welche Ursachen und Konsequenzen von Kundenorientierung Inhalt der bisherigen wissenschaftlichen Auseinandersetzung mit dem Thema waren.

Der erste Ansatz, Kundenorientierung als wissenschaftliches Konstrukt zu fassen, stammt von Saxe und Weitz (1982, S. 343), die Kundenorientierung als „the degree to which salespeople practice the marketing concept by trying to help their customers make purchase decisions that will satisfy customer needs“ beschreiben und damit die Unterstützung der Kundschaft und deren Zufriedenheit mit der getroffenen Kaufentscheidung in den Mittelpunkt ihrer Definition stellen. Anhand einer Literaturrecherche und anschließender Experteninterviews entwickelten sie einen Item-Pool, aus dem sie nach einer ersten Befragung die besten 24 Items auswählten, um die SOCO (Selling Orientation – Customer Orientation) -Skala zu bilden. Die faktorielle Struktur der Items, die Reliabilität der Skala sowie ein korrelativer Zusammenhang mit der Verkaufsleistung wurden empirisch bestätigt.

Michaels und Day (1985) argumentieren, dass die Einschätzung der Kundenorientierung durch die Kundschaft möglicherweise objektiver ist als die Selbsteinschätzung durch Verkaufsmitarbeiter. Sie passen die Formulierung der Items der SOCO Skala so an, dass sie eingesetzt werden konnte, um die Kundenorientierung des Verkaufspersonals durch die Kundschaft einschätzen zu lassen. Ihre Ergebnisse zeigen eine leicht höhere Reliabilität der angepassten Skala und eine sehr ähnliche faktorielle Struktur wie bereits von Saxe und Weitz (1982) berichtet. Deutliche Unterschiede zeigten sich in der Einschätzung, wie hoch die Kundenorientierung ausgeprägt ist – die Einschätzung durch die Kundschaft fiel deutlich geringer aus als die Selbsteinschätzung des Verkaufspersonals. Eine Erklärung für diesen Unterschied könnte eine positive Selbstdarstellung oder Überschätzung der Kundenorientierung durch das Verkaufspersonal, eine Unterschätzung der Kundenorientierung durch die Kundschaft oder eine Mischung beider Effekte liefern. Die Arbeit von Michaels und Day (1985) macht deutlich, dass das Konstrukt Kundenorientierung aus mindestens den beiden Perspektiven verkaufende und kaufende Person betrachtet werden kann.

Die Perspektive der Klienten bzw. Einkäufer wurde auch von Tadepalli (1995) untersucht. Er leitete aus Interviews mit Einkäufern ab, dass die Items von Michaels und Day (1985) so angepasst werden sollten, dass diese sich auf die letzte Kaufsituation und eine spezifische Verkaufsperson beziehen. Zudem passte er das bislang neunstufige Antwortformat auf ein siebenstufiges Antwortformat an. Die Unidimensionalität und eine hohe Reliabilität der resultierenden Items konnten empirisch bestätigt werden. Auch die Konvergente-, Diskriminante- und Konstruktvalidität konnte in einer Untersuchung durch konfirmatorische Faktorenanalysen bestätigt werden (Tadepalli, 1995).

Brown et al. (2002) konzipieren Kundenorientierung als Persönlichkeitsfacette und definieren Kundenorientierung als „an employee's tendency or predisposition to meet customer needs in an on-the-job context“. Sie gehen davon aus, dass Kundenorientierung die beiden Dimensionen needs und enjoyment umfasst. Die Subdimension need bildet die Annahme des Personals ab, inwiefern sie die Bedürfnisse der Kundschaft befriedigen können. Die Facette enjoyment

beschreibt, ob Arbeitskräfte Gefallen an der Interaktion mit der Kundschaft und deren Bedienung finden. Sie entwickelten einen Fragebogen, der jede dieser Subdimensionen mit sechs Items abbildet, und konnten faktorenanalytisch zeigen, dass sich die beiden theoretisch angenommenen Subskalen empirisch bestätigen lassen. Die Reliabilität der Gesamtskala sowie der Subskalen fiel hoch aus. Die hohe Interkorrelation der beiden Subskalen, rechtfertigt die Annahme, dass diese Subfacetten des übergeordneten Konstruktes Kundenorientierung sind.

Um die inkrementelle Validität des Konstruktes Kundenorientierung über etablierte Persönlichkeitsfacetten, wie sie im weit verbreiteten Fünf-Faktoren Modell der Persönlichkeitspsychologie angenommen werden, bei der Vorhersage von Arbeitsleistungsindikatoren zu quantifizieren, untersuchen Brown et al. (2002), ob die Vorhersage der Arbeitsleistung durch gängige Persönlichkeitskonstrukte partiell oder vollständig durch Kundenorientierung mediiert wird. Abbildung 4 zeigt die von ihnen getesteten Modelle.

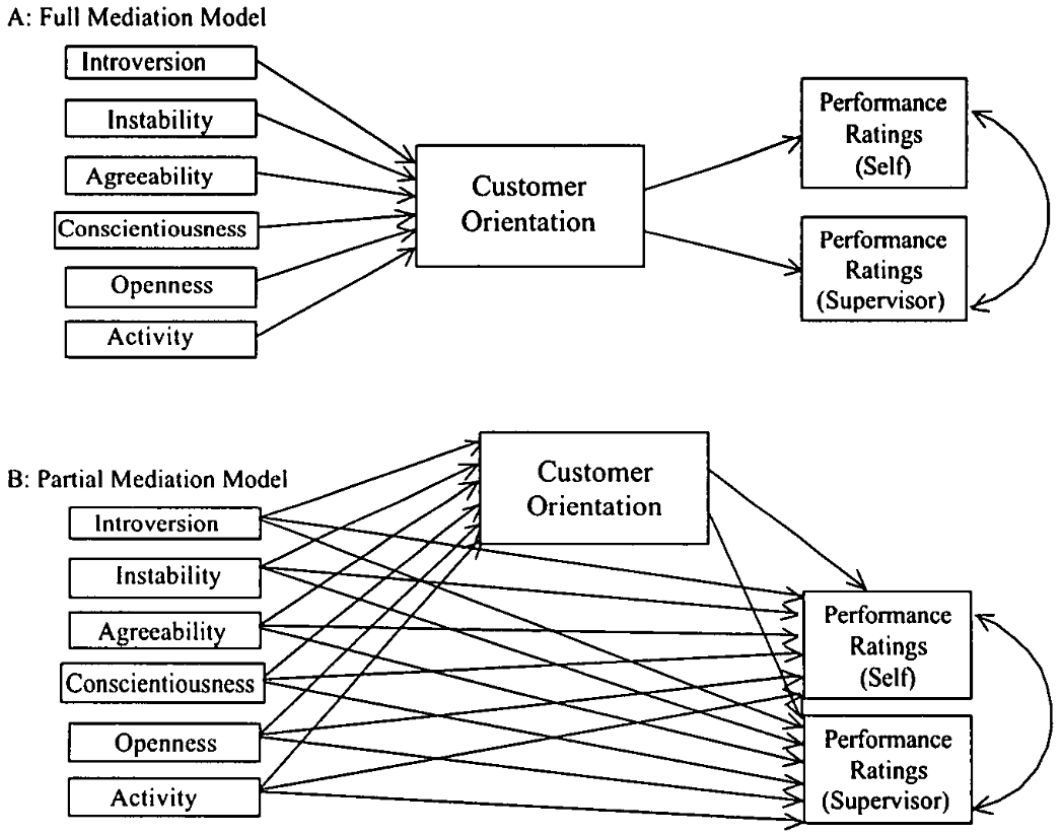


Abbildung 4 Rolle der Kundenorientierung bei der Vorhersage von Arbeitsleistung (Brown et al., 2002)

Die Ergebnisse ihrer Studie zeigen, dass beide in Abbildung 4 dargestellten Modelle einen guten Modellfit aufweisen, wobei das in Teil B der Abbildung dargestellte Modell signifikant besser zu

den erhobenen Daten passt. Dass die unter den Annahmen von Modell B geschätzte Varianz-Kovarianz-Matrix besser zu den empirischen Daten passt, ist unabhängig von theoretischen Erklärungen auf die Tatsache zurückzuführen, dass dieses Modell weniger restriktiv ist. Im gewählten Strukturgleichungsmodell konnte 26 Prozent der Varianz der selbst eingeschätzten Arbeitsleistung und 12 Prozent der Varianz der durch Vorgesetzte eingeschätzten Arbeitsleistung aufgeklärt werden. Diese Ergebnisse zeigen, dass das Konstrukt Kundenorientierung als Persönlichkeitskonstrukt auf einem höheren Abstraktionsniveau einen Beitrag zur Vorhersage der Arbeitsleistung leisten kann, der über den der klassischen Persönlichkeitseigenschaften aus dem Fünf-Faktoren Modell hinausgeht. Aus den Ergebnissen lässt sich weiterhin erkennen, dass sich die Kundenorientierung eines Mitarbeiters am besten durch die Persönlichkeitskonstrukte „Agreeability", „Activity" und „Instability" vorhersagen lässt. Hohe Kundenorientierung geht mit hoher Verträglichkeit einher und zeigt sich bei Personen, die altruistisch handeln, anderen wohlwollend, hilfsbereit und mit Verständnis und Mitgefühl begegnen. Diese Personen kooperieren gerne, sind nachgiebig und vertrauen anderen. Der gefundene positive Zusammenhang zwischen Kundenorientierung und dem Konstrukt „Activity" legt nahe, dass Personen, die das Bedürfnis haben, körperlich aktiv zu sein, hohe Kundenorientierung zeigen. Kundenorientierung steht in einem negativen Zusammenhang mit emotionaler Instabilität; Personen, die ängstlich, nervös, angespannt, unsicher und verlegen sind, weisen demnach eine geringere Kundenorientierung auf (Brown et al., 2002).

Welche weiteren Einflüsse Kundenorientierung mit sich bringt, untersuchen Donavan et al. (2004). Sie modellieren ihre Annahmen als latentes Strukturgleichungsmodell und stellen fest, dass Kundenorientierung in positivem Zusammenhang mit Arbeitszufriedenheit und Organisationsverbundenheit steht. Diese Zusammenhänge fielen hypothesenkonform für Personen, die mehr Zeit im direkten Kundenkontakt verbrachten, höher aus. Als vermittelnde Variable konnten sie die, durch das Personal selbst eingeschätzte, Passung zu den Anforderungen der Arbeitsstelle identifizieren. Ihre Ergebnisse verdeutlichen, dass für Personen im direkten Kundenkontakt Kundenorientierung, verstanden als Persönlichkeitseigenschaft, eine wichtige Voraussetzung ist, um sich selbst geeignet für die Tätigkeit und zufrieden mit der Arbeit im Kundenkontakt zu erleben.

Ob Kundenorientierung ein psychologisches Konstrukt ist, das sich auf die arbeitsrelevanten Variablen Stressempfinden und Engagement auswirkt, oder ob es sich bei Kundenorientierung eher um Verhaltensweisen des Personals handelt, die durch Variablen des Arbeitsumfeldes verursacht werden, wird von Zablah, Franke, Brown und Bartholomew (2012) im Rahmen einer Metaanalyse untersucht. Sie wählten 291 Publikationen aus und fassten deren Ergebnisse zu einer Korrelationsmatrix zusammen, die sie als Ausgangspunkt für Strukturgleichungsmodelle nutzten, die die verschiedenen konkurrierenden theoretischen Konzeptionen abbilden. Ihre Ergebnisse unterstützen die Sichtweise, dass Kundenorientierung Stresserleben und Engagement im Arbeitsumfeld beeinflusst und dass diese Variablen Teil des Mediationsmechanismus sind,

der hinter dem Zusammenhang zwischen Kundenorientierung und der Arbeitsleistung besteht. Rod und Ashill (2010) integrieren das Konstrukt Kundenorientierung als Prädiktorvariable in ein Modell (siehe Abbildung 5), in dem als emotionale Auswirkungen Arbeitszufriedenheit und Organisationsverbundenheit und als arbeitsbezogene Auswirkungen die Serviceleistung und die Absicht, die aktuelle Position im Unternehmen zu verlassen, vorhergesagt wurden.

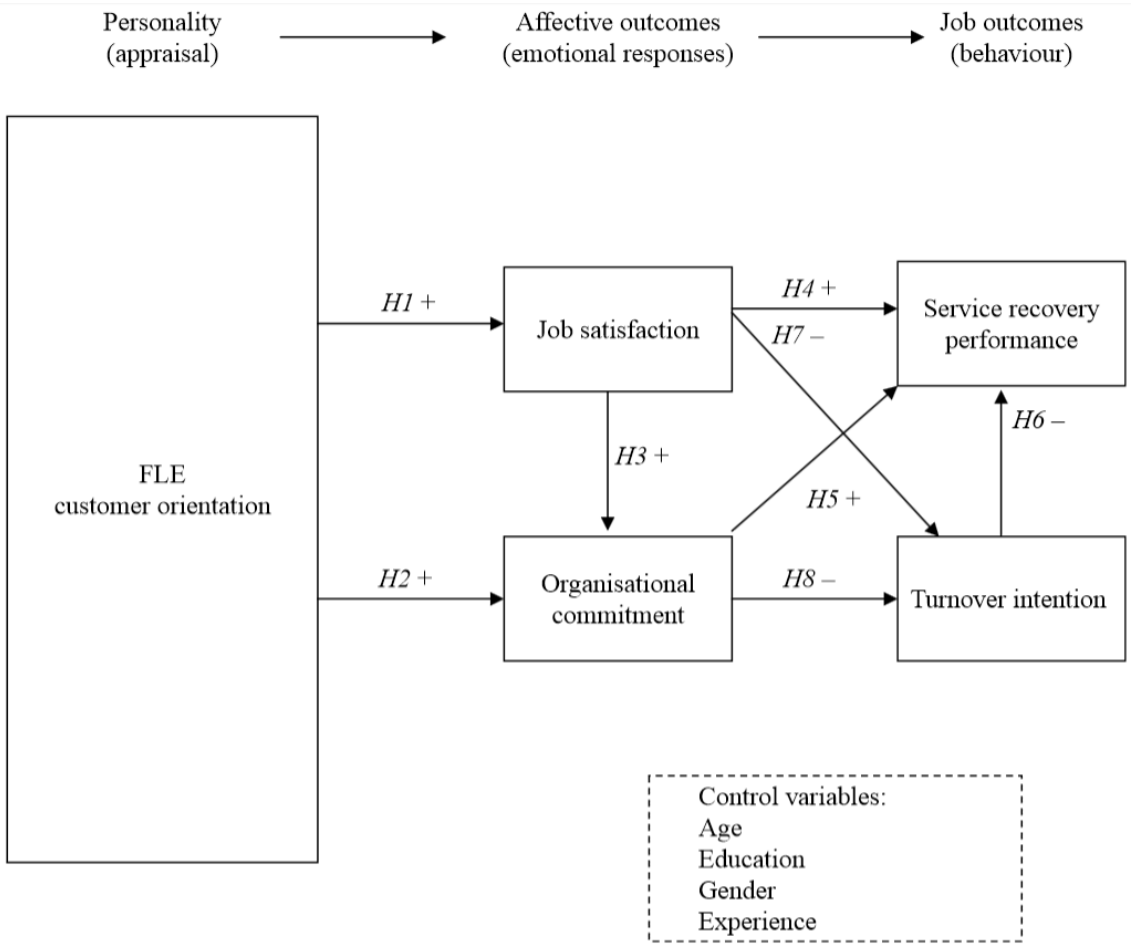


Abbildung 5 Auswirkungen von Kundenorientierung aus (Rod & Ashill, 2010)

Sie modellierten ihre Hypothesen als Strukturgleichungsmodell und konnten ihre Annahmen weitestgehend bestätigen. Lediglich ihre Hypothese 8 (siehe Abbildung 5), dass Organisationsverbundenheit Einfluss auf die Fluktuation des Personals hat, ließ sich nicht bestätigen.

Von welchen Bedingungen Kundenorientierung in Produktionsbetrieben abhängig ist, wird von Liao und Subramony (2008) untersucht, indem sie Fragebogendaten von Teammitgliedern und Führungskräften, aus 42 Produktionsstätten in 16 Ländern mit einem hierarchischen Mehrebenmodell analysierten. Sie unterscheiden drei Rollen innerhalb einer Organisation

(Personen im direkten Kundenkontakt, Arbeitskräfte in der Produktion und Personal, das unterstützende Funktion hat, wie zum Beispiel im Personalwesen oder in der Buchhaltung), die unterschiedlich intensiven Kontakt mit externen Kunden haben und nahmen an, dass die Kundenorientierung der Personen höher ausgeprägt ist, wenn diese in engem Kontakt zu Kunden stehen. Zudem vermuteten sie einen Zusammenhang zwischen der Kundenorientierung der Führungskräfte und der Kundenorientierung der Teammitglieder. Ihre Analyseergebnisse bestätigen die theoretischen Annahmen und können insofern generalisiert werden, als sie in unterschiedlichen Produktionsstätten, ansässig in unterschiedlichen Ländern, nachgewiesen wurden. Das gewählte Untersuchungsdesign lässt jedoch keine kausalen Schlüsse zu und die Frage, warum Personen, die mehr Kundenkontakt haben, eine höhere Kundenorientierung aufweisen, konnte deshalb nicht geklärt werden.

Im Fokus einer Studie von Homburg et al. (2011) stehen die Rahmenbedingungen, die den Zusammenhang zwischen Kundenorientierung und Kundenbindung moderieren. Sie unterscheiden zwischen funktionaler Kundenorientierung, die aus aufgabenorientierten Verhaltensweisen, wie zum Beispiel der Beschreibung von Produkten oder der Identifikation der Bedürfnisse der Kundschaft, besteht und relationaler Kundenorientierung, bei der der Aufbau einer persönlichen Beziehung zur Kundin oder zum Kunden im Vordergrund steht. Abbildung 6 zeigt, wie sie den moderierenden Einfluss des Kommunikationsstils und der Eigenschaften des angebotenen Produkts konzipieren.

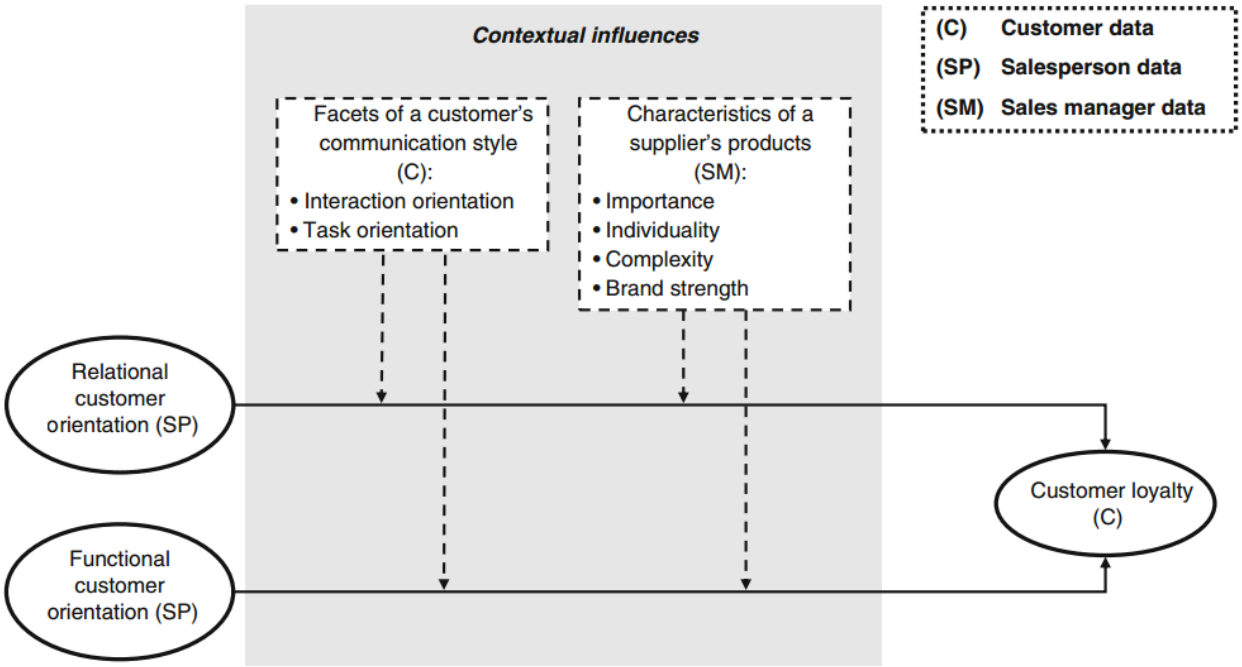


Abbildung 6 Moderierte Effekte von Kundenorientierung auf Kundenbindung aus (Homburg et al., 2011)

In Abbildung 6 wird rechts oben dargestellt, dass die erfassten Variablen sich auf hierarchisch geschachtelten Ebenen manifestieren und aus Befragungen der Kundschaft, des Verkaufspersonals und der Managerinnen und Manager stammen. Die Datenanalyse mittels Mehrebenenanalyse zeige, dass Kundenbindung durch funktionale Kundenorientierung, aber nicht durch relationale Kundenorientierung vorhergesagt werden kann. Funktionale Kundenorientierung erweist sich, insbesondere bei Produkten, die hohe Wichtigkeit aufweisen, als guter Prädiktor für Kundenbindung. Zudem lässt sich die Loyalität der Kundschaft durch deren Interesse, eine starke persönliche Beziehung zum Verkaufspersonal aufzubauen, vorhersagen. Die Studie von Homburg et al. (2011) zeigt insgesamt, dass die Annahme, dass Kundenorientierung grundsätzlich zu hoher Kundenbindung führt, nicht unter allen Umständen gültig ist, und liefert wichtige Hinweise darauf, dass weitere Variablen, die sich möglicherweise auf einer anderen Betrachtungsebene manifestieren, in die Theoriebildung einfließen sollten.

Rafaeli, Ziklik und Doucet (2008) nutzen transkribierte Telefongespräche von Angestellten eines Bankhauses und leiten mittels qualitativer Verfahren fünf Kategorien von kundenorientierten Verhaltensweisen ab. Basierend auf diesem Kategoriensystem kodieren sie die untersuchten Gespräche und zeigen, dass es einen Zusammenhang gibt zwischen der Häufigkeit, mit der kundenorientierte Verhaltensweisen gezeigt wurden, und der durch die Kundschaft eingeschätzten Service-Qualität. Mit ihrer Studie liefern Rafaeli et al. (2008) wertvolle Hinweise, welche Subfacetten das Konstrukt Kundenorientierung aufweist, und zeigen auf, dass die Einschätzung der Service-Qualität durch die Kundschaft in Zusammenhang mit der Kundenorientierung der Angestellten steht.

Die Zusammenfassung der mittlerweile zahlreich vorliegenden Forschungsarbeiten zum Thema Kundenorientierung machen deutlich, dass in der Frage, ob Kundenorientierung eher als ein Bündel von Verhaltensweisen verstanden oder als domänenspezifisches Persönlichkeitskonstrukt konzipiert werden sollte, noch keine Einigkeit besteht. In der Literatur finden sich zahlreiche Ansätze, Kundenorientierung in Form von relativ unterschiedlichen Fragebogen-Items zu operationalisieren. Hinter den unterschiedlichen Operationalisierungen stehen unterschiedliche Auffassungen darüber, welche Subfacetten von Kundenorientierung angenommen werden. Auch die Frage, ob die Kundenorientierung am besten durch das Verkaufspersonal selbst eingeschätzt werden kann oder ob diese durch die Kundschaft oder Vorgesetzte eingeschätzt werden sollte, ist noch nicht abschließend geklärt. Die vorliegenden Studien zeigen einen Zusammenhang zwischen Kundenorientierung und Arbeitsleistung des Personals, wobei verschiedene Variablen wie zum Beispiel Stressempfinden und Arbeitszufriedenheit als Mediator- oder Moderatorvariablen identifiziert werden. Auch der Zusammenhang zu dem in dieser Arbeit zentralen Konstrukt Service-Qualität konnte empirisch gezeigt werden.

3.2.5 Kundenzufriedenheit

Hohe Service-Qualität und Kundenzufriedenheit stellen zentrale Ziele der meisten Organisationen dar. Sie werden häufig in einem Atemzug genannt und als zwei Seiten einer Medaille verstanden. In diesem Kapitel wird dargestellt, warum es für eine wissenschaftliche Betrachtung sinnvoll und wichtig ist, diese beiden Konstrukte zu unterscheiden, und welche Bedingungen und Konsequenzen von Kundenzufriedenheit durch die Forschung bislang identifiziert wurden.

Ein theoretisches Modell zur Kundenzufriedenheit von Oliver (1980) wird von Rust und Oliver (1994, S. 2) folgendermaßen beschrieben:

„In brief, customer satisfaction is a summary cognitive and affective reaction to a service incident (or sometimes to a long-term service relationship). Satisfaction (or dissatisfaction) results from experiencing a service quality encounter and comparing that encounter with what was expected." Bereits in dieser frühen Definition wird Kundenzufriedenheit als kognitive und emotionale Reaktion beschrieben, die bei einem Service-Ereignis aus dem Vergleich zwischen den Erwartungen und dem Erlebten entsteht.

Nach Braun und Haferburg (2001, S. 13) ist Kundenzufriedenheit „der emotionale Zustand und die damit verbundene kognitive Bewertung hinsichtlich eines Produkts oder einer Dienstleistung". Diese Definition verdeutlicht, dass Kundenzufriedenheit sowohl eine emotionale als auch eine kognitive Komponente umfasst. Andere Definitionen, wie beispielsweise von Rapp (1997, S. 6): „Kundenzufriedenheit ist eine individuelle Einstellung, die durch den permanenten Vergleich der tatsächliche wahrgenommenen Unternehmensleistung und den Erwartungen bezüglich dieser Unternehmensleistung entsteht", beschreiben das Konstrukt als Einstellung und rücken den Vergleich zwischen der Unternehmensleistung und den Erwartungen der Kundschaft in den Vordergrund.

Kundenzufriedenheit kann an der Schnittstelle zwischen den Aktivitäten von Organisationen und deren Kunden angesiedelt werden. Die Aktivität der Organisation besteht in der Regel darin, Produkte und/oder Dienstleistungen anzubieten. Auf der Seite der Kundenaktivitäten interessieren potenzielle Beschwerden, die Bereitschaft das Angebot weiter zu empfehlen, die Rückkehr der Kundschaft und besonders deren Absicht, weiterhin loyal zu bleiben, da diese unmittelbar mit dem Umsatz einer Organisation verknüpft ist (Homburg, 2011).

Das bekannteste Modell, um Kundenzufriedenheit und ihre Entstehung zu beschreiben, stellt das Confirmation/Disconfirmation-Paradigma (C/D-Paradigma) dar, in dem davon ausgegangen wird, dass Kundenzufriedenheit aus dem Vergleich von Erlebnissen bei der Nutzung von Produkten oder Dienstleistungen (Ist-Leistung) und Vergleichsstandards oder Erwartungen der Kundschaft (Soll-Leistung) resultiert (Oliver & Swan, 1989). Entsprechen bei diesem Vergleich Soll- und Ist-Leistung einander, fällt die Kundenzufriedenheit auf Konfirmationsniveau aus. Wird die Ist-Leistung besser als erwartet eingeschätzt, spricht man von positiver Diskonfirmation, die zu Kundenzufriedenheit über dem Konfirmationsniveau führt. Als negative

Diskonfirmation bezeichnet man Situationen, in denen das Erlebnis mit dem Produkt oder der Dienstleistung schlechter als erwartet ausfällt, was zu Kundenzufriedenheit unterhalb des Konfirmationsniveaus führt.

Anderson und Sullivan (1993) testen die Annahmen des C/D-Paradigmas und kommen erwartungskonform zu dem Ergebnis, dass sich Kundenzufriedenheit aus der Abweichung aus Erwartungen und Ist-Leistung vorhersagen lässt. Sie stellen fest, dass es einen positiven Zusammenhang zwischen den Erwartungen und der Einschätzung der Ist-Leistung gibt und dass Erwartungen und die wahrgenommene Ist-Leistung als einzelne Prädiktoren Kundenzufriedenheit ebenfalls vorhersagen können.

Die Grundidee, Zufriedenheit als Soll-Ist-Diskrepanz zu erklären, wurde bereits von Locke (1969) kritisiert, der darauf hinweist, dass Abweichungen von Erwartungen von Individuen für unterschiedlich wichtig erachtet werden und somit unterschiedlich gewichteten Einfluss auf die Zufriedenheitseinschätzung haben. Zudem geht er auf die Schwierigkeit ein, dass für die Intensität der Zufriedenheit, die Erwartungen, die Einschätzung der Ist-Leistung und die individuelle Wichtigkeit der Diskrepanz keine einheitliche physikalische oder psychologische Maßeinheit existiert. Das Berechnen der Zufriedenheit als Soll-Ist Diskrepanz, die durch die Wichtigkeit gewichtet wird, erscheint vor diesem Hintergrund fragwürdig.

Kanning und Bergmann (2009) vergleichen in einer empirischen Studie das einfache C/D-Paradigma mit einer erweiterten Variante, die die Idee von Locke (1969) aufgreift und die Wichtigkeit der Diskrepanz einbezieht. Sie stellen fest, dass das erweiterte Modell keine signifikant bessere Vorhersage der Kundenzufriedenheit ermöglicht als das einfach C/D Paradigma. Sie schließen aus ihren Ergebnissen, dass die zusätzliche Variable, die die subjektive Wichtigkeit der Diskrepanz abbildet, für die Vorhersage der Kundenzufriedenheit von geringer Nützlichkeit ist.

Kundenzufriedenheit kann auch durch die Anwendung der Equity Theorie, die sich mit dem Thema Gerechtigkeit befasst, erklärt werden (Adams, 1963). Personen erkaufen eine Leistung oder ein Produkt zu einem bestimmten Preis, dabei schätzen sie ein, ob das Verhältnis aus Preis und erhaltener Gegenleistung für sie angemessen ausfällt. Wird dieses Verhältnis als ausgeglichen empfunden, wird der Austausch nach der Equity Theorie als gerecht empfunden, was sich positiv auf die Kundenzufriedenheit auswirken sollte (Szymanski & Henard, 2001).

Die Metaanalyse von Szymanski und Henard (2001) repliziert den Befund von Anderson und Sullivan (1993) und weist ebenfalls darauf hin, dass die Ist-Leistung, also die wahrgenommenen Eigenschaften des Produkts oder der Dienstleistung, auch unabhängig von bestehenden Erwartungen einen direkten Effekt auf die Einschätzung der Kundenzufriedenheit hat.

Wie bereits in der Definition des Konstruktes von Braun und Haferburg (2001) angedeutet, ist Kundenzufriedenheit auch mit emotionalen Zuständen verknüpft. Wirtz und Bateson (1999) erweitern das C/D-Paradigma, das einen rein kognitiven Ansatz darstellt, indem sie die Rolle von Emotionen integrierten. In ihrer Studie belegten sie empirisch, dass positive Diskonfirma-

tion mit mehr Freude und negative Diskonfirmation mit weniger Freude einhergeht. Für die Emotion Freude konnten sie einen direkten Effekt auf die Kundenzufriedenheit nachweisen. Der Einfluss von positiven Emotionen auf Kundenzufriedenheit kann auch in der Metaanalyse von Szymanski und Henard (2001) wiedergefunden und damit als empirisch gesichert betrachtet werden.

Ein Modell, dass sowohl die Entstehung und Konsequenzen von Kundenzufriedenheit umfasst, stammt von Braun und Müssigmann (2009a) und wird in Abbildung 7 dargestellt.

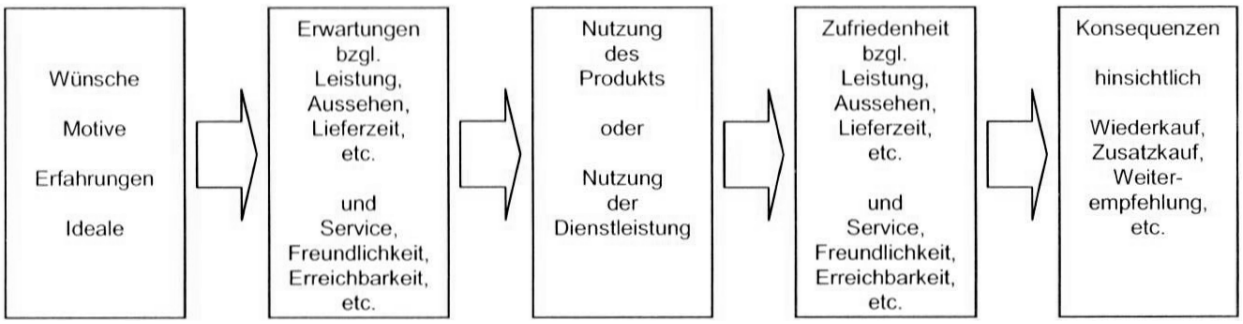


Abbildung 7 Modell der Entstehung und Konsequenzen von Kundenzufriedenheit aus Braun und Müssigmann (2009a)

Auch in diesem Modell finden sich die Elemente Erwartungen und die Nutzungserfahrungen aus dem C/D-Paradigma wieder. Es verdeutlicht, dass Kundenzufriedenheit auf Produkte und Dienstleistungen bezogen werden kann, und zeigt, dass hohe Kundenzufriedenheit zu Wiederkauf, Zusatzkauf und Weiterempfehlung führt. Zudem weisen Braun und Müssigmann (2009a, S. 11) darauf hin, dass „es nicht nur die objektiven Produkteigenschaften sind, die zur Kundenzufriedenheit führen, sondern dass auch die Qualität des Service einen bedeutenden Beitrag zur Entstehung von Kundenzufriedenheit leistet".

Fischer, Braun, Kehr und Schreiber (2012) untersuchen, wie die beiden Konstrukte Kundenzufriedenheit und Service-Qualität zusammenhängen. Sie zeigen in einer empirischen Studie, dass Service-Qualität einen direkten Einfluss auf Kundenzufriedenheit hat, zudem arbeiten sie heraus, dass sich hohe Service-Qualität auf das Image der Organisation auswirkt und das Image der Organisation ebenfalls Einfluss auf die Kundenzufriedenheit hat.

Einen mathematischen Ansatz, um den Wert von Kundenzufriedenheit auszudrücken, entwickeln Rust und Zahorik (1993). Sie gehen davon aus, dass Kundenzufriedenheit zu Kundenbindung führt und sich damit positiv auf den Marktanteil und den Umsatz eines Unternehmens auswirkt. Ihr mathematisches Modell kann genutzt werden, um zu berechnen, wie viele finanzielle Ressourcen man zur Verbesserung der Kundenzufriedenheit einsetzten sollte.

Rapp (1997) entwickelt ein Modell, in dem technische Produktqualität, Service-Qualität, Reputationsqualität, persönliche Beziehungsqualität und die Preiswahrnehmung als Prädiktoren zur Vorhersage der Kundenzufriedenheit berücksichtigt werden. Weiterhin wird in seinem Modell angenommen, dass Kundenzufriedenheit mit der Loyalität der Kundschaft verknüpft ist. Eine empirische Analyse seines Modells zeigt, dass der wichtigste Prädiktor für Kundenzufriedenheit die Service-Qualität darstellt; zudem bestätigt sich deutlich der Zusammenhang zwischen der Zufriedenheit und der Loyalität der Kundschaft.

Es haben sich verschiedene Strategien durchgesetzt, um Kundenzufriedenheit zu messen. Zum einen können, wie im C/D-Paradigma gefordert, die Einschätzung der Ist-Leistung und Erwartungen erfragt werden, um anschließend aus der Differenz dieser Werte auf die Kundenzufriedenheit zu schließen. Eine andere Operationalisierungsstrategie besteht darin, Personen im Rahmen der Kundenzufriedenheitsbefragung anzuleiten, die von ihnen wahrgenommene Ist-Leistung mit ihren Erwartungen abzugleichen und anschließend direkt nach der Zufriedenheit oder danach zu fragen, ob die Erwartungen erfüllt, über-, oder unterschritten wurden. Ein weiterer Ansatz, der in der Kundenzufriedenheitsforschung häufig genutzt wird, besteht darin, direkt nach der Zufriedenheit und möglichen Folgeaktivitäten, wie Beschwerden, der Bereitschaft das Produkt oder die Dienstleistung weiterzuempfehlen und der Absicht, erneute bzw. zusätzliche Käufe zu tätigen, zu fragen (Fürst, 2011). Bei standardisierten Befragungen mit geschlossenem Antwortformat ist in der Kundenzufriedenheitsforschung häufig festzustellen, dass fast ausschließlich von denjenigen Antwortkategorien Gebrauch gemacht wird, die für sehr hohe Kundenzufriedenheit stehen. Dieser Deckeneffekt in den Befragungsdaten geht damit einher, dass die Varianz im Antwortverhalten eingeschränkt ist und deshalb schlecht zwischen sehr zufriedenen und extrem zufriedenen Kunden differenziert werden kann. Um in diesem Bereich hoher Kundenzufriedenheit gut differenzieren zu können, müssen Items eingesetzt werden, die eine sehr hohe Item-Schwierigkeit aufweisen. Es lohnt sich, auch das genutzte Antwortformat zu überarbeiten, so dass weniger Antwortkategorien für den Bereich geringer Kundenzufriedenheit angeboten und Antwortalternativen hinzugefügt werden, die extrem hohe Kundenzufriedenheit ausdrücken (Klarmann, 2011).

Dieses Phänomen lässt sich damit erklären, dass insbesondere Stammkundschaft in der Regel mit dem getätigten Handel zufrieden ist, da sie sich sonst für konkurrierende Angebote entscheiden würden. Eine positive Einschätzung des zurückliegenden Handelns dient darüber hinaus dazu, kognitive Dissonanz zu vermeiden. Neben standardisierten Befragungen, die als Fragebogen auf Papier oder als Onlineumfrage umgesetzt werden können, kann auch telefonisch befragt werden. Ein anderer Weg, an Informationen zur Kundenzufriedenheit und Verbesserungsmöglichkeiten zu gelangen, schlagen Braun und Müssigmann (2009b) durch Gruppendiskussionen mit der Kundschaft vor.

Weitere Forschungsfragen der Kundenzufriedenheitsforschung, die nicht unmittelbar mit dem Thema dieser Arbeit verknüpft sind, befassen sich damit, wie Erwartungen bzw. Vergleichsstandards geformt werden, wie die Wahrnehmung und Einschätzung der Ist-Leistung abläuft und wie sich Kundenzufriedenheit von dem Einstellungskonstrukt abgrenzen lässt.

Will man aus inhaltlichen und theoretischen Gründen zwischen Kundenzufriedenheit und Service-Qualität unterscheiden, wird deutlich, dass Kundenzufriedenheit, wie in diesem Kapitel beschrieben, in der Auseinandersetzung mit Angeboten von Organisationen auf der Seite der Kundschaft entsteht und eingeschätzt wird. Service-Qualität beschreibt, welche Eigenschaften die angebotenen Dienstleistungen haben. Die Tatsache, dass beide Konstrukte im Alltag vermengt werden, ist darauf zurückzuführen, dass es einen klaren Zusammenhang zwischen diesen Konstrukten gibt, der so interpretiert werden kann, dass hohe Service-Qualität eine Voraussetzung für hohe Kundenzufriedenheit ist.

3.3 Das Konstrukt Service-Qualität

Das Konstrukt Service-Qualität, das in der englischsprachigen Fachliteratur unter dem Stichwort „service quality“ auftaucht, wird in deutschsprachigen Publikationen auch mit dem synonym verwendeten Begriff Dienstleistungsqualität bezeichnet. Der Begriff Service, mit der Bedeutung Dienstleistung, Kundendienst, Kunden- oder Gästebedienung, wurde in der ersten Hälfte des 20. Jahrhunderts vom englischen Wort service ins Deutsche übernommen. Etymologisch geht das englische Wort service auf das altfranzösische servise zurück, das wiederum seinen Ursprung im lateinischen Wort servitium hat, das für Sklavendienst, Sklaverei oder übertragen jede Art von Dienstbarkeit steht (Pfeifer, 1989).

3.3.1 Besonderheiten und Abgrenzung des Konstruktes

Die in Kapitel 3.2.1 beschriebenen Konzepte, die sich auf Produkte beziehen, können auf den Bereich von Service bzw. Dienstleistungen nicht ohne weiteres übertragen werden. Dies wird besonders deutlich, wenn man sich einige zentrale Unterschiede zwischen Dienstleistungen und Produkten vor Augen führt. Schneider et al. (1997) weisen darauf hin, dass sich Dienstleistungen von Produkten in drei Dimensionen unterscheiden: „intangibility“ (Greifbarkeit), „simultaneity“ (Simultanität oder Gleichzeitigkeit) und „customer participation in production“ (Partizipation des Kunden an der Herstellung). Ein zentraler Unterschied zwischen Waren und Dienstleistungen kann in der Dimension Greifbarkeit gesehen werden. Je ungreifbarer ein Angebot ist, desto mehr handelt es sich bei diesem um einen Service. Ein gutes Beispiel, um dies zu verdeutlichen, ist ein Konzert. Verglichen mit einem Produkt kann ein Konzertbesuch nicht direkt angefasst werden, es gibt nichts, was eingepackt und mit nach Hause genommen

und auch künftig genutzt werden könnte. Eine weiteres Merkmal, das Dienstleistungen von Waren unterscheidet, ist nach Schneider et al. (1997) die Simultanität. Während Waren häufig an einem Ort produziert, gelagert und zu einem bestimmten Zeitpunkt ausgeliefert werden, werden Dienstleistungen meistens zeitgleich hergestellt, geliefert und konsumiert. Dies macht es in der Regel unmöglich, analog zur Produktion von Waren über Qualitätskontrollen sicherzustellen, dass ein Service den gesetzten Standards entspricht, bevor die Kundschaft damit in Kontakt kommt. Diese Sonderstellung von Service führt dazu, dass Service-Personal beim Erbringen von Dienstleistungen bereits beim ersten Kundenkontakt und grundsätzlich spontan alles richtig machen müssen. Zudem muss die gewünschte Dienstleistung zu dem Zeitpunkt erbracht werden, zu dem sie benötigt wird, und kann im Normalfall nicht vollständig vorbereitet oder gelagert werden. Die dritte von Schneider et al. (1997) identifizierte Dimension betrifft die Tatsache, dass die Kundschaft bei Dienstleistungen meist an der Erstellung der Dienstleistung beteiligt ist. Während die Produktion von Waren meist zeitlich und räumlich getrennt von den kaufenden oder nutzenden Personen erfolgt, setzen Dienstleistungen, wie z. B. der Transport durch öffentliche Verkehrsmittel, häufig voraus, dass die Kundschaft anwesend ist und sich zum Beispiel durch das Kaufen einer Fahrkarte oder das Ein-, Um- und Aussteigen am Erbringen der Dienstleistung beteiligt. Die meisten Angebote befinden sich nicht an den Extremen der angeführten Dimension und bei genauer Betrachtung ist häufig festzustellen, dass die meisten Dienstleistungen durch dazugehörige Produkte ergänzt werden und fast zu allen Produkten begleitend Serviceleistungen angeboten werden.

3.3.2 Perspektiven und Zugänge zum Konstrukt Service-Qualität

Der Blick auf das Thema Service-Qualität kann aus verschiedenen Perspektiven erfolgen. Geht es darum, Service-Qualität zu erfassen und zu steuern, richtet sich der Fokus von Organisationen häufig auf die eigene Service-Qualität. Um diese zu steuern, werden vom Management häufig Vorgaben oder Ziele gesetzt, die ähnlich einer Norm einen Soll-Wert definieren. Die Kontrolle, ob solche Vorgaben tatsächlich umgesetzt werden, kann im Rahmen einer managementorientierten Erfassung von Service-Qualität durch Controlling oder Benchmarking umgesetzt werden. Ergänzt wird diese Strategie häufig durch eine mitarbeiterorientierte Messung, bei der das Verhalten der Organisationsmitglieder untersucht wird (Urban, 2013). In Verhaltensbeobachtungen, Einzelinterviews, Gruppendiskussionen oder mit Hilfe von standardisierten Fragebögen wird erfasst, welches Wissen das Personal über servicebezogene Zielvorgaben hat und welche servicerelevanten Verhaltensweisen gezeigt werden. Da solche Versuche, Service-Qualität zu messen, meistens reaktiv sind, also einen Einfluss auf die betroffenen Personen und das zu erfassende Verhalten haben, kann deren Messergebnis absichtlich oder unbewusst verfälscht sein (Adair, 1984; Olson, Verley, Santos & Salas, 2004). Um ein möglichst unverzerrtes Abbild der tatsäch-

lichen Service-Qualität zu bekommen, bietet es sich an, Indikatoren für Service-Prozesse, wie zum Beispiel Wartezeiten, objektiv zu erfassen. Solche Indikatoren sind häufig mittels technischer Lösungen relativ einfach erfassbar und sind nicht durch Darstellungstendenzen oder sozial erwünschtes Antwortverhalten beeinflusst.

Neben der Perspektive der Organisation wird Service-Qualität sehr häufig aus der Sicht der Kundschaft betrachtet. Die von Kundinnen und Kunden wahrgenommene Qualität einer Dienstleistung hängt stark mit der tatsächlichen Güte der Dienstleistung zusammen, bildet diese jedoch nicht zwangsläufig perfekt ab. Am Beispiel von Produktqualität zeigen Mitra und Golder (2006), dass objektive Veränderungen der Qualität von der Kundschaft zeitversetzt wahrgenommen werden. Zudem wirken sich Faktoren wie die Art der Qualitätsveränderung, das Markenimage und Produktspezifische Variablen auf die Wahrnehmung der Qualität durch die Kundschaft aus. Dagger und Sweeney (2007) weisen darauf hin, dass die Wahrnehmung von Service-Qualität davon beeinflusst wird, wie lange das Kundenverhältnis bereits besteht, und dass spezifische Elemente von Service-Qualität von Neukunden anders eingeschätzt werden als von langjährigen Bestandskunden. Auch das Wissen über Service-Konzepte der Organisation hat Einfluss auf die Einschätzung der Service-Qualität und das Vertrauen der Kundschaft (Eisingerich & Bell, 2008). Um bei einer kundenorientierten Messung von Service-Qualität möglichst hohe Objektivität zu ermöglichen, werden Service-Indikatoren definiert und unter anderem durch das sogenannte „Mystery Shopping“ eingeschätzt (Finn & Kayande, 1999; Lazarus, 2009; Minghetti & Celotto, 2013). Bei dieser Herangehensweise geht es darum, dass geschulte Personen beauftragt werden, verdeckt in Interaktion mit den Produkten und Dienstleistungen einer Organisation in Kontakt zu treten und dabei gemäß vorgegebener Instruktionen zu beurteilen, wie gut bestimmte Elemente des Service-Prozesses umgesetzt werden. Da diese Herangehensweise sehr zeit- und kostenaufwendig ist, werden in vielen Organisationen Kundenbefragungen eingesetzt, um eine subjektive Einschätzung der Kundschaft zu erhalten (Braun & Müssigmann, 2009b). Solche Kundenbefragungen basieren sehr häufig auf einer Kombination aus standardisierten Fragebögen mit geschlossenem Antwortformat und offenen Fragen, über die eine freie Rückmeldung gewonnen werden kann. Um die auf diesem Weg gewonnenen Rückmeldungen zu validieren und zu vertiefen, können in Interviews oder Gruppendiskussionen relevante Themenfelder genauer analysiert werden.

Oft rückt bei Kundenbefragungen das Thema Kundenzufriedenheit in den Vordergrund, das, wie in Kapitel 3.2.5 beschrieben, eine Konsequenz von ausreichend hoher Service-Qualität sein kann. Die Einschätzung der Service-Qualität kann durch verschiedene Faktoren verzerrt werden. Personen, die erhebliche Kosten und großen Aufwand aufgewendet haben, neigen zur Rechtfertigung ihres Verhaltens und damit zu positiven Einschätzungen der Service-Qualität (Cardozo, 1965).

Da der Kundschaft der Blick hinter die Kulissen einer Organisation meist verwehrt ist, kann sie nur Teile der Service-Qualität einschätzen. Organisationsinterne Prozesse, die maßgeblich an hoher Service-Qualität beteiligt sind, können von der Kundschaft nicht beurteilt werden. Zudem achten Personen häufig selektiv auf bestimmte Aspekte des Kundenkontakts, die dann bei der Einschätzung der Service-Qualität andere Elemente überstrahlen und somit deren Einschätzung verzerren können (van Doorn, 2008).

Wenn Organisationen sich mit Service-Qualität auseinandersetzten, entwickeln sie häufig Befragungsinstrumente, deren Items sehr spezifisch auf Service-Leistungen der eigenen Organisation bzw. Branche zugeschnitten sind. Solche individuell angepassten Ansätze liefern sehr konkrete Hinweise, in welchen organisationsspezifischen Bereichen Verbesserungspotenziale existieren, haben jedoch den Nachteil, dass die Ergebnisse nicht mit anderen Organisationen und Branchen verglichen werden können. Um die so erfasste Service-Qualität einordnen zu können, müssten organisations- oder branchenspezifische Referenzwerte generiert werden. Derartige Instrumente können darüber hinaus aufgrund der spezifischen Inhalte häufig nur in der Organisation sinnvoll eingesetzt werden, für die sie entwickelt wurden.

In der wissenschaftlichen Auseinandersetzung mit Service-Qualität ist die Generalisierbarkeit der Ergebnisse auf andere Unternehmen und Branchen häufig von großer Bedeutung. Dies setzt voraus, dass die entwickelten Instrumente in unterschiedlichen Organisationen eingesetzt werden können, weshalb meist allgemeine Aussagen zum Thema Service-Qualität genutzt werden. Die Ergebnisse derartige Instrumente können für den Vergleich zwischen verschiedenen Organisationen und Branchen herangezogen werden, liefern jedoch häufig eher allgemeine Ansatzpunkte, die für Verbesserungsmaßnahmen im Kontext der im Fokus stehenden Organisation, im Dialog mit den Organisationsmitgliedern, spezifiziert werden müssen.

3.3.3 Definitionen

Um das Konstrukt für die akademische Auseinandersetzung greifbar, abgrenzbar und messbar zu machen, ist es wichtig zu definieren, was genau unter Service-Qualität verstanden wird. Parasuraman, Zeithaml und Berry (1985) entwickeln bereits 1983 erste Ideen und Konzepte zum Thema Service-Qualität, die sie zu dem in Abbildung 8 dargestellten Lückenmodell integrieren. Von der Kundschaft wahrgenommene Service-Qualität definieren sie „als das Ausmaß der Diskrepanz zwischen den Erwartungen und Wünschen der Kundschaft und ihren Eindrücken von der tatsächlichen Leistung“ (Zeithaml, Parasuraman & Berry, 1992, S. 32).

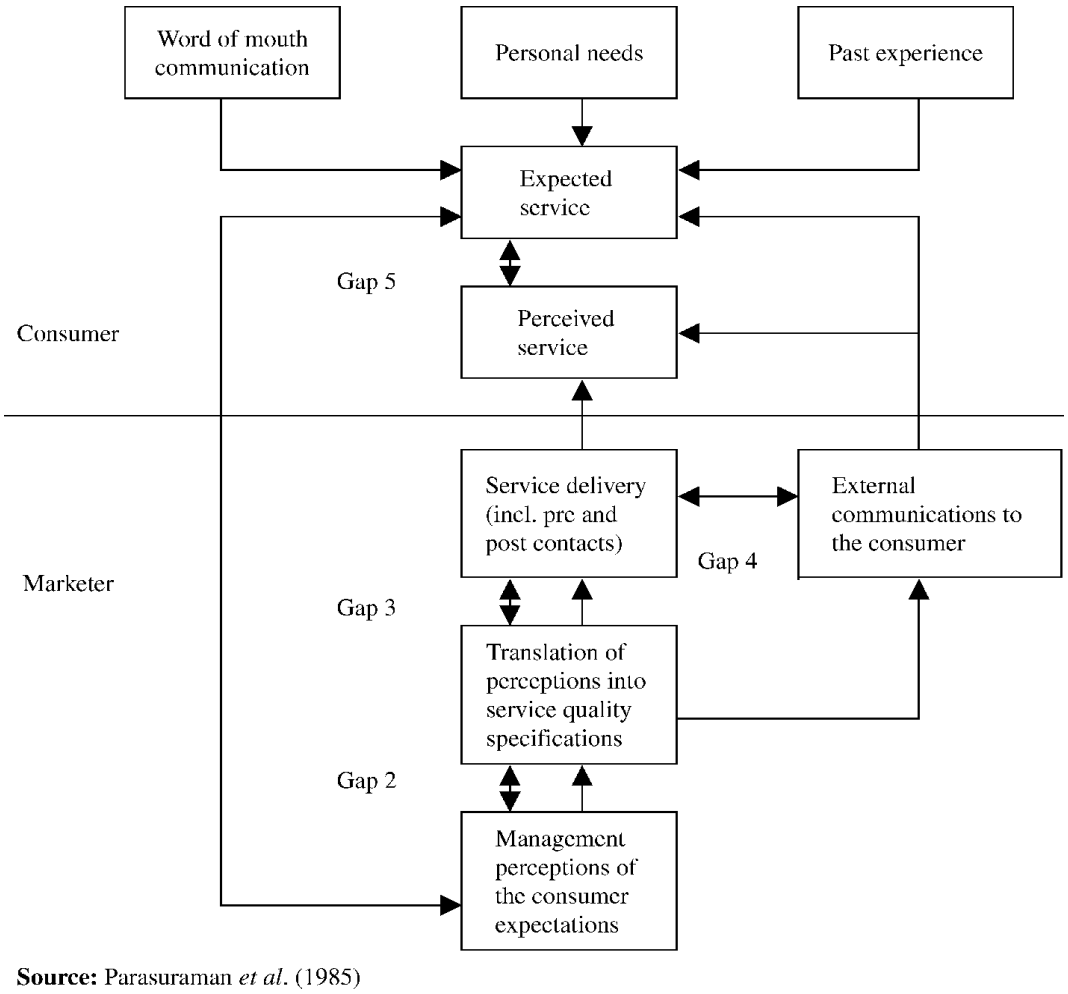


Abbildung 8 Lückenmodell nach Parasuraman et al. (1985)

Diese Definition und der Ansatz des Lückenmodells basieren, ähnlich wie das C/D-Paradigma der Kundenzufriedenheitsforschung, das in Kapitel 3.2.5 beschrieben wurde, auf der Differenz zwischen Erwartungen und der wahrgenommenen Qualität der Interaktion mit den Servicemitarbeitern bzw. den Produkten einer Organisation. Das Lückenmodell beschreibt darüber hinaus, welche Faktoren sich auf die Service Erwartungen der Kundschaft auswirken und wie das Management einer Organisation solche Service-Erwartungen durch externe Unternehmenskommunikation beeinflussen kann. Zudem weist das Modell darauf hin, dass Organisationen diese Service-Erwartungen als Grundlage für interne Service-Standards und -Prozesse nutzen können.

Ein weiteres Modell, das mehr oder weniger zeitgleich von dem finnischen Marketingforscher Christian Grönroos vorgeschlagen wurde, beschreibt Service-Qualität ebenfalls als Konstrukt, das zwischen dem erwarteten und dem wahrgenommenen Service steht (Grönroos, 1984). Zudem betont dieser als „nordisches Modell“ in die Literatur eingegangene Ansatz, wie wichtig

das Image der Gesamtorganisation oder die lokale Niederlassung bei der Einschätzung der wahrgenommenen Service-Qualität ist.

Auch Rust und Oliver (1994) fassen zusammen: „We conclude that perceived service quality is a subjective matter…“ (S. 10) und betonen damit, dass die Service-Qualität identischer Dienstleistungen oder Produkte von Person zu Person unterschiedlich wahrgenommen und eingeschätzt wird. Zudem definieren sie drei Hauptelemente von Service-Qualität: Service Product, Service Environment und Service Delivery.

Bitner und Hubbert (1994) definieren Service-Qualität als „The consumer's overall impression of the relative inferiority / superiority of the organization and its services“ (S.77). Sie beschreiben Service-Qualität als ein Konstrukt höherer Ordnung, dem ein höheres Abstraktionsniveau zugrunde liegt und grenzen Service-Qualität damit von der Zufriedenheit mit einzelnen Service-Kontakten und der Gesamtzufriedenheit mit allen Service-Interaktionen ab.

Eine aktuellere Definition bietet Bruhn (2010) an, der Service-Qualität bzw. Dienstleistungsqualität als „die Fähigkeit eines Anbieters, die Beschaffenheit einer primär intangiblen und der Kundenbeteiligung bedürfenden Leistung gemäß den Kundenerwartungen auf einem bestimmten Anforderungsniveau zu erstellen“ (S. 38), beschreibt. Auch in dieser Definition spielt die Berücksichtigung der Kundenerwartungen eine zentrale Rolle

Diese exemplarisch dargestellten Definitionen des Konstrukts zeigen, dass im Großteil der Forschung zu Service-Qualität Kundinnen und Kunden und deren Einschätzungen im Fokus stehen. Diese Kundenperspektive ist nur ein Zugang zum Thema Service-Qualität, der in zwei Punkten wesentlich kritisiert werden kann. Das Urteil der Kundschaft ist sehr subjektiv und von deren Erwartungen geprägt. Zudem hat die Kundschaft in der Regel keinen direkten Einblick in die Service-Prozesse, die innerhalb einer Organisation entwickelt und gelebt werden. Sich bei der Erfassung von Service-Qualität nur auf Kundenurteile zu verlassen, erscheint daher zweifelhaft. Zudem stellt sich die Frage, welche und wie viele Kundengruppen für eine zuverlässige Einschätzung der Service-Qualität befragt werden müssen. Selbst wenn eine Organisation sich für die zeit- und kostenintensive Strategie entscheidet, alle Kundinnen und Kunden zu diesem Thema zu befragen, werden bestimmte Kundengruppen, zum Beispiel extrem unzufriedene Personen, sich an solchen Befragungen nicht beteiligen. Somit wird deutlich, dass trotz hohen Aufwands auf diesem Weg nur bedingt zuverlässige Informationen über die tatsächliche Service-Qualität einer Organisation gewonnen werden können.

3.3.4 Etablierte Operationalisierungen des Konstrukts

Einer der ersten Ansätze zur Erfassung von Service-Qualität, wahrgenommen aus der Sicht der Kundschaft, stammt von Parasuraman, Zeithaml und Berry (1988). Ihr Ausgangspunkt, zehn, sich potenziell überlappende Dimensionen, die sie aus Aussagen ableiteten, die bei der

Beschreibung von Service-Qualität genutzt wurden. Sie formulierten 97 Items zu diesen Dimensionen und erfassten jeweils die generelle Erwartung an Organisationen und die spezifische Einschätzung des erlebten Service im Kontakt mit einer spezifischen Organisation, auf einer siebenstufigen Antwortskala. Die Differenzwerte zwischen dem wahrgenommenen Service und dem erwarteten Service bildeten die Grundlage zur faktorenanalytischen Untersuchung und Bestimmung der Reliabilität der Skalen mittels Cronbach's α. Diese empirische Analyse führte zur Überarbeitung des Instruments, die erneut empirisch validiert wurde, woraus eine Version mit 22 Item-Paaren entstand, die den in Tabelle 1 dargestellten Dimensionen zugeordnet wurden.

Tabelle 1 Dimensionen des SERVQUAL aus Parasuraman et al. (1988)

Dimension	**Kurzbeschreibung**
Tangibles	Physical facilities, equipment, and appearance of personnel
Reliability	Ability to perform the promised service dependably and accurately
Responsiveness	Willingness to help customers and provide prompt service
Assurance	Knowledge and courtesy of employees and their ability to inspire trust and confidence
Empathy	Caring, individualized attention the firm provides its customers

Die so entwickelte SERVQUAL-Skala zur Erfassung von Service-Qualität wurde in zahlreichen Studien weiterentwickelt und findet auch in aktuellen Studien Beachtung (Parasuraman et al., 1991b; Parasuraman, Berry & Zeithaml, 1991a; Berry & Parasuraman, 1992; Brown, Churchill Jr & Peter, 1993; Parasuraman, Berry & Zeithaml, 1993; Parasuraman, 1994; Parasuraman, Zeithaml & Berry, 1994; Ali & Raza, 2017).

Die Operationalisierung von Service-Qualität anhand der Diskrepanz zwischen erwartetem und wahrgenommenem Service, wie sie im Lückenmodell (siehe Abbildung 8) definiert ist, wird aus verschiedenen Gründen kritisiert (Cronin Jr & Taylor, 1992; Teas, 1993). Diese Autoren favorisieren eine leistungsbasierte Erfassung des Konstrukts als Einstellung. Ihr Forschungsansatz führte zur SERVPERF-Skala, die durch den Verzicht, für jedes Item die Erwartung abzufragen, halb so lang ausfällt (Cronin Jr & Taylor, 1992; Cronin Jr & Taylor, 1994; Barros Jerônimo & Medeiros, 2014). Die höhere Sparsamkeit und Ökonomie des Verfahrens, die psychometrischen Eigenschaften der Items und Skalen sowie ihre empirischen Befunde zum Zusammenhang von Service-Qualität, Kundenzufriedenheit und Kaufabsicht weisen auf die Überlegenheit der SERVPERF-Skala hin (Cronin Jr & Taylor, 1994; Jain & Gupta, 2004). Zudem argumentieren Cronin Jr und Taylor (1994), dass die Ergebnisse der SERVPERF-Skala dem Management einer Organisation einen guten Gesamtüberblick geben kann, der sich für

einen Vergleich mit anderen Organisationen sowie für eine längsschnittliche Darstellung, die auch subgruppenspezifisch sein kann, gut eignet. Der Vergleich von SERVQUAL- und SERVPERF-Skala von Rodrigues, Barkur, Varambally und Golrooy Motlagh (2011) zeigt, dass die beiden Instrumente, im Bildungssektor eingesetzt, zu unterschiedlichen Einschätzungen der Service-Qualität führen. Auch das Korrelationsmuster mit externen Validitätskriterien unterscheidet sich, so dass die Autoren zur Empfehlung kommen, beide Instrumente zeitgleich zu nutzen.

Um die Akzeptanz der SERVQUAL- bzw. SERVPERF-Skala in der praktischen Anwendung in Organisationen zu steigern, haben zahlreiche Autoren Varianten entwickelt, die auf ihren spezifischen Anwendungsbereich zugeschnitten sind. Vaughan und Shiu (2001) arbeiteten in Kommunalverwaltungen, im Wohltätigkeitssektor und im ehrenamtlichen Bereich. Sie entwickelten die ARCHSECRET Skala, die aus qualitativen Fokusgruppeninterviews mit ihrer Kundschaft abgeleitet wurde und die etablierten Dimensionen der SERVQUAL-Skala um sechs domänenspezifische Dimensionen erweitert. Wie wichtig menschliche Faktoren und Benutzerfreundlichkeit im Bereich der Service-Qualität sind, unterstreichen Strawderman und Koubek (2008) und erweitern die SERVQUAL-Skala zur SERVUSE-Skala, die neben den klassischen Dimensionen auch die Facette Benutzerfreundlichkeit umfasst.

Für den Bereich der elektronischen Dienstleistungen (e-service), der zum Beispiel Internetshops umfasst und den Rust (2001, S. 283) als „the provision of service over electronic networks" definiert, wurden zahlreiche Operationalisierungen von Service-Qualität entwickelt. Fassnacht und Koese (2006) geben einen Überblick über diese Ansätze und fassen sie zu einem konzeptionellen Modell mit neun Faktoren erster Ordnung zusammen. Sie entwickelten ein Befragungsinstrument, um ihr Modell zu operationalisieren und konnten anhand konfirmatorischer Faktorenanalysen zeigen, dass die empirischen Befunde ihre angenommene Modellstruktur bestätigen. Parasuraman, Zeithaml und Malhotra (2015) entwickelten die E-S-Qual-Skala und die E-RecSQUAL. Diese Skalen sind auf die spezifischen Service-Elemente des Internetversandhandels zugeschnitten, wurden empirisch überprüft und weisen gute psychometrische Eigenschaften auf.

Shemwell und Yavas (1999) entwickelten ein domänenspezifisches Modell zur Messung von Service-Qualität in Krankenhäusern. Ihr Modell enthält die drei Dimensionen „Search", „Experience" und „Credence". Die Subdimension „Search", die mit fünf Items erfasst wird, umfasst Eigenschaften, über die nahezu alles erfahren werden kann, bevor die Kaufentscheidung getroffen wird. Eigenschaften eines Produktes oder einer Dienstleistung, die erst nach dem Kauf und der Inanspruchnahme der Dienstleistung rational bewertet werden können, sind in der Subdimension „Experience" mit fünf Items erfasst. Die Subdimension „Credence", die vier Items enthält, umfasst Eigenschaften, bei denen nicht genug Wissen verfügbar ist, um eine vernünftige Einschätzung vorzunehmen. Diese Gliederung des Konstrukts Service-Qua-

lität in drei Subdimensionen erfolgt theoriegeleitet und wird mittels verschiedener Modelle der konfirmatorischen Faktorenanalyse bestätigt.

Einen guten Überblick über verschiedene Maße für Service-Qualität liefert Ladhari (2008), der 30 einschlägige Studien zusammenfasst und darstellt, welche Maße jeweils für Service-Qualität genutzt wurden. Er weist darauf hin, dass neben der SERVQUAL Skala mittlerweile zahlreiche branchenspezifische Skalen existieren, die Wissenschaftler und Praktiker Hilfe und Orientierung beim Entwickeln und Einsetzen von branchenspezifischen Instrumenten für Service-Qualität bieten.

3.3.5 Stand der Forschung zum Thema Service-Qualität

Ein Blick in die publizierte Literatur zum Thema Service-Qualität erweckt leicht den Eindruck, dass deutlich mehr Arbeiten zu den Konsequenzen von Service-Qualität vorliegen als zu den Voraussetzungen und Bedingungen, die zu mehr oder weniger Service-Qualität führen. Aus theoretischer Perspektive wird deutlich, dass jede Arbeit, die sich mit der Definition oder Operationalisierung des Konstruktes befasst, sehr präzise Annahmen darüber macht, welche Elemente das Konstrukt umfasst. Welche derartigen Elemente als Teil des Konstruktes beschrieben werden oder ob angenommen wird, dass es sich bei spezifischen Elementen um Bedingungen bzw. Voraussetzungen für Service-Qualität handelt, sind Fragen der wissenschaftlichen Theoriebildung. In einer Studie von Dabholkar, Shepherd und Thorpe (2000) wird versucht, diese Frage empirisch zu klären. Ihre Ergebnisse zeigen, dass ein Strukturgleichungsmodell, das gängige Subdimensionen von Service-Qualität nicht als Komponenten, sondern als Bedingungen für Service-Qualität auffasst, bessere Passung zu den erhobenen Daten aufweist. Auch Wang, Lo und Hui (2003) zeigen, dass Operationalisierungen der fünf Service Dimensionen des SERVQUAL empirisch gut als Prädiktoren für eine insgesamt eingeschätzte Service-Qualität bestätigt werden können. Berücksichtigt man diese Ausführungen, so kann man zu der Auffassung kommen, dass es mindestens genauso viel, wenn nicht sogar deutlich mehr Publikationen gibt, die sich mit der Abgrenzung des Konstrukts und damit auch mit den Bedingungen und Voraussetzungen von Service-Qualität befassen. Neben den in Kapitel 3.3.3 und 3.3.4 beschriebenen Facetten von Service-Qualität unterstreichen Gounaris, Stathakopoulos und Athanassopoulos (2003) die Wichtigkeit einer persönlichen Beziehung zwischen dem Service-Personal und der Kundschaft sowie einer erwartungsformenden Weitergabe von Informationen und Empfehlungen in persönlichen Gesprächen im sozialen Umfeld der Kundschaft. Auch die Vertrautheit mit der Marke und deren Dienstleistungen sowie die Einschätzung darüber, wie sehr sich Anbieter an den individuellen Bedingungen des Marktes orientieren, können als Bedingungen für das Zustandekommen von hoher Service-Qualität gesehen werden (Gounaris et al., 2003).

Einer der am häufigsten untersuchten und empirisch am besten abgesicherten Zusammenhänge besteht zwischen Service-Qualität und dem Erfolg von Organisationen (Borucki & Burke, 1999; Ryan, Schmit & Johnson, 1996; Schneider et al., 1998; Schulte, Ostroff, Shmulyian & Kinicki, 2009; Taylor & Baker, 1994; Zohar, 2000). Hui et al. (2007, S. 151) gehen mit ihrer Aussage „...the success of a business largely depends on the quality of service provided to its customers" sogar so weit zu behaupten, dass der Erfolg einer Organisation größtenteils von der Service-Qualität abhängt. Ogden und Watson (1999) zeigen an einem Beispiel aus der privatisierten Wasserwirtschaft, dass die Analyse empirischer Daten mittels Mehrebenenmodell aufdeckte, wie eine Verbesserung im Kundenservice zu mehr Umsatz und Rendite für die Aktionäre führen. Schneider et al. (1997, S. 49) fassen den Stand ihrer Erkenntnis wie folgt zusammen: „Enough evidence is now accumulating to permit the following conclusion: Service quality can yield customer loyalty, and customer loyalty reduces the cost of selling new services; then the wonders of compounding take over to yield organizational success." Sie nehmen also empirisch begründet an, dass Service-Qualität im Zusammenhang mit Kundenbindung steht. Weiterhin behaupten sie, dass Kundebindung Kosten reduziert und damit im Zusammenhang mit dem Erfolg von Organisationen steht. Sie formulieren damit eine Mediationshypothese, die den bivariaten Zusammenhang zwischen Service-Qualität und dem Erfolg einer Organisation über die Kundenbindung vermittelt modelliert.

Taylor und Baker (1994) untersuchten, wie sich die Kaufabsicht durch Service-Qualität und Kundenzufriedenheit vorhersagen lässt. Sie gehen davon aus, dass der Zusammenhang zwischen Service-Qualität und Kaufabsicht durch Kundenzufriedenheit moderiert wird. In ihren Regressionsanalysen zeigen sich hypothesenkonform signifikante Effekte der Interaktion aus Service-Qualität und Kundenzufriedenheit, wobei die β-Koeffizienten des Interaktionsterms relativ niedrig ausfallen. In einer Studie von Brady und Robertson (2001) wurde ebenfalls das Ziel verfolgt, das Kundenverhalten durch Kundenzufriedenheit und Service-Qualität vorherzusagen. Die Autoren testen zwei konkurrierende Mediationsmodelle; in einem wurde der Effekt von Service-Qualität auf Kundenverhalten durch Kundenzufriedenheit mediiert, im anderen der Effekt von Kundenzufriedenheit auf das Kundenverhalten durch Service-Qualität. Ihre Ergebnisse zeigen über verschiedene Kulturen hinweg, dass die Annahme, Kundenzufriedenheit mediiere den Zusammenhang zwischen Service-Qualität und Kundenverhalten, empirisch untermauert werden kann. Auch im Bereich Luftfahrt können Etemad-Sajadi, Way und Bohrer (2016) in ihrer Studie bestätigen, dass Service-Qualität einen direkten Effekt auf Kundenzufriedenheit und Kundenloyalität hat. Neben diesen direkten Effekten zeigt sich ebenfalls ein indirekter Effekt von Service-Qualität auf Kundenloyalität, der über Kundenzufriedenheit mediiert wird.

Harrison-Walker (2016) untersuchte, welche Rolle Service-Qualität bei der Verbreitung von Empfehlungen und Informationen durch mündliche Weitergabe im persönlichen Gespräch spielt. Er stellt heraus, dass im Bereich der Veterinärindustrie bei geringer Service-Qualität

Mund-zu-Mund-Kommunikation häufiger und detaillierter erfolgt, und vermutet, dass dahinter das Ziel steckt, andere vor potenziellen Problemen mit Dienstleistern zu warnen.

Zahlreiche wissenschaftliche Publikationen integrieren die Erkenntnisse über Bedingungen und Konsequenzen von Service-Qualität in komplexere Modelle. Ein zentrales Modell, das die Zusammenhänge zwischen Service-Qualität, Kundenzufriedenheit und dem Erfolg von Organisationen beschreibt, ist unter dem Namen Service-Profit Chain bzw. im deutschen Sprachraum unter dem Namen Wert-Gewinnkette-Modell bekannt geworden (Heskett, Jones, Loveman, Sasser & Schlesinger, 1994; Heskett, Sasser & Schlesinger, 2015). Abbildung 9 zeigt, wie Heskett et al. (1994) die Elemente der Service-Profi Chain und deren Zusammenhänge beschreiben.

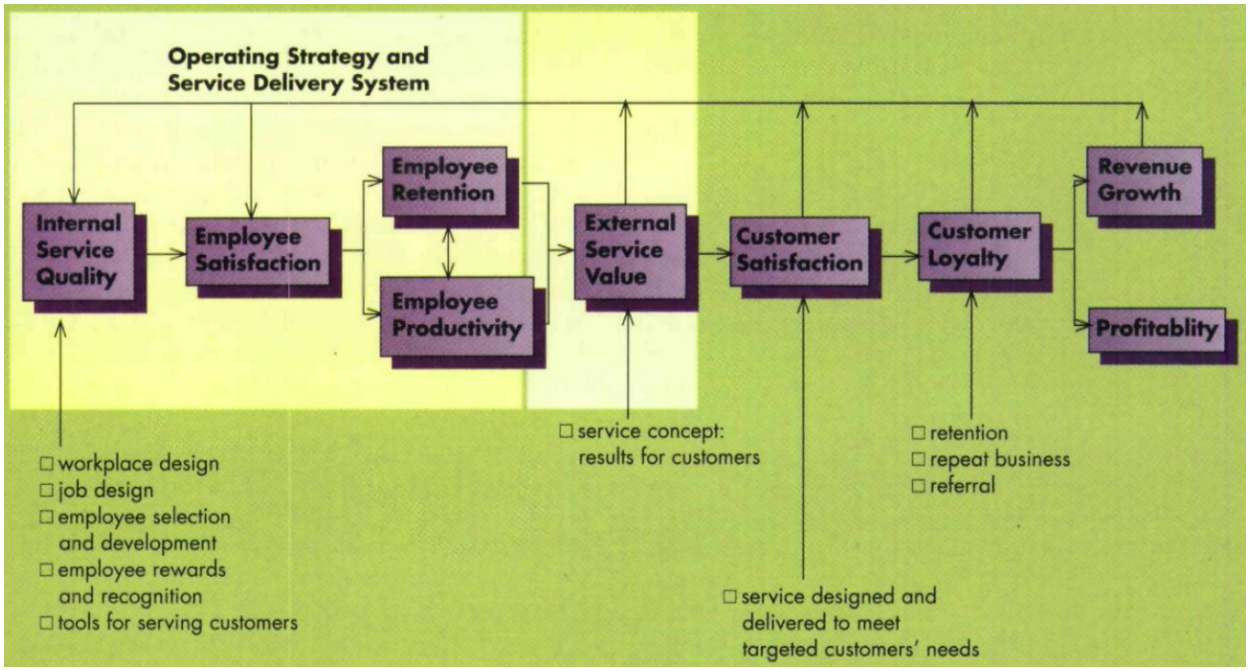


Abbildung 9 Zusammenhänge innerhalb der Service-Profit Chain, aus Heskett et al. (1994)

Heskett et al. (1997) nehmen an, dass der Zusammenhang zwischen Service-Qualität und dem Erfolg von Organisationen als Mediationsmodell beschrieben werden kann. Sie identifizieren unter anderem Mitarbeiter- und Kundenzufriedenheit als vermittelnde Variablen zwischen der Service-Qualität und dem Erfolg der Organisation. Ihr Modell enthält keinen direkten Einfluss der Service-Qualität auf den Erfolg der Organisation. Für den beobachtbaren bivariate Zusammenhang zwischen Service-Qualität und dem Erfolg der Organisation, werden als vermittelnden Variablen Kundenzufriedenheit, Kundenbindung und Mitarbeiterzufriedenheit angeführt. Zudem nehmen sie an, dass der Erfolg der Organisation und die Kundenzufriedenheit Auswirkungen auf die Service-Qualität und die Mitarbeiterzufriedenheit haben. Heskett (2014)

erweitert sein Modell zum „culture impact model“, indem er die Mitarbeiterzufriedenheit und Mitarbeiterbindung als weitere Bedingungen für höhere Produktivität und mehr Profit in die Grundstruktur der Service-Profit Chain integrierte.

Die Grundidee der Service-Profit Chain wurde in der Forschung mehrfach aufgegriffen und bestätigt (Heskett, Sasser & Schlesinger, 2003; Hogreve, Iseke, Derfuss & Eller, 2017; Homburg, Wieseke & Hoyer, 2009; Hong et al., 2013; Kamakura, Mittal, Rosa & Mazzon, 2002; Maxham, Netemeyer & Lichtenstein, 2008).

Gelade und Young (2005) untersuchen im Bankensektor, wie Service-Klima, Kundenzufriedenheit und Verkaufsleistung zusammenhängen (Abbildung 10), und stellen fest, dass die empirisch gefundenen Mediationseffekte zwar signifikant, jedoch in geringer Höhe ausfallen.

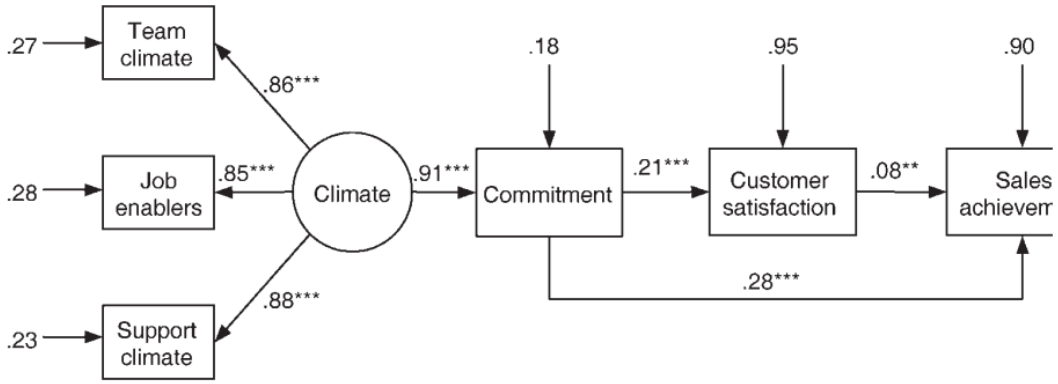


Abbildung 10 Service-Profit Chain aus Gelade und Young (2005)

Schneider, Ehrhart, Mayer, Saltz und Niles-Jolly (2005) entwickeln ein ähnliches Mediationsmodell, in dem sich Service-Qualität auf Kundenzufriedenheit und Kundenzufriedenheit auf Verkaufszahlen auswirkt. Wie in Abbildung 11 dargestellt, beginnt für sie Service-Qualität mit entsprechendem Führungsverhalten und lässt sich als Service-Klima beschreiben, das zu kundenfokussiertem Verhalten führt.

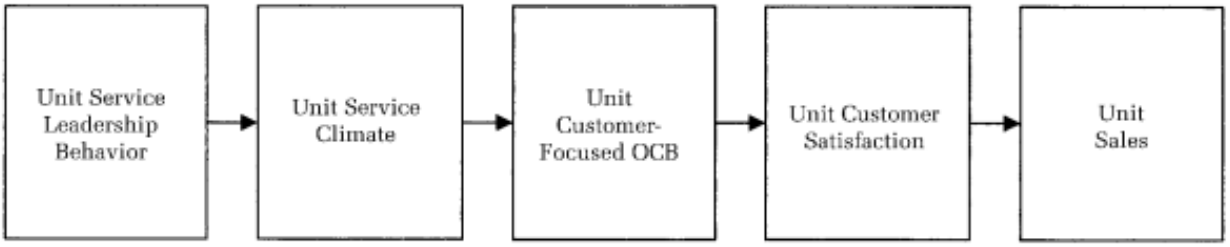


Abbildung 11 Service-Profit Chain aus Schneider et al. (2005)

Eine zentrale Fragestellung in ihrer Studie ist es, empirisch zu untersuchen, ob die Annahme einer vollständigen Mediation, wie in Abbildung 11 dargestellt, plausibel ist, oder ob direkte Effekte der Prädiktorvariablen zu beobachten sind. Der Vergleich der Fit-Statistiken für die verschiedenen konkurrierenden Strukturgleichungsmodelle sprach für die Annahme eines reinen Mediationsmodells.

Um Studien, die das Modell der Service-Profit Chain untersuchen, zusammenzufassen, führen Hong et al. (2013) eine Metaanalyse durch, in der sie 58 Primärstudien berücksichtigen. Sie leiten aus den vorliegenden Befunden ein theoretisches Modell ab, das in Abbildung 12 dargestellt ist.

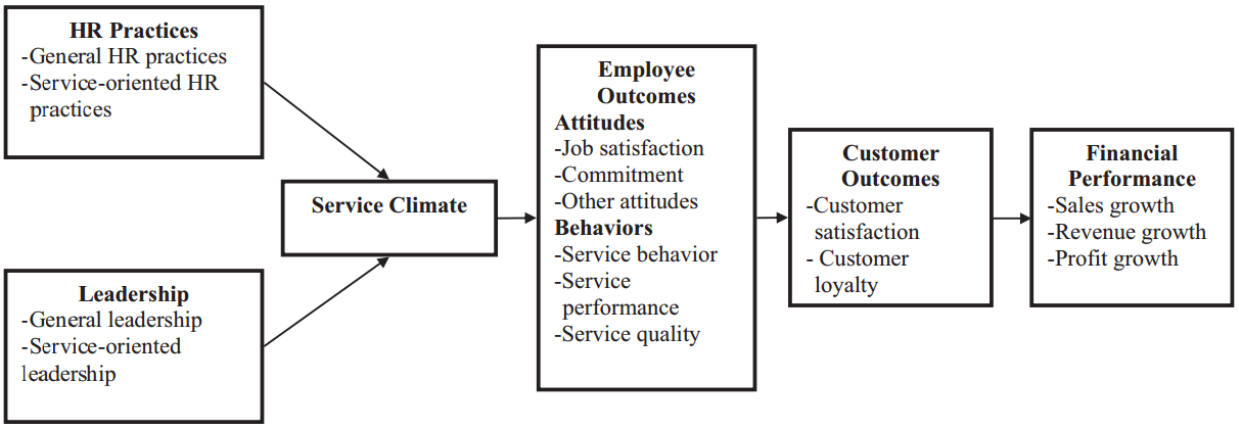


Abbildung 12 Theoretisches Modell von (Hong et al., 2013)

Service-Qualität wird im Modell von Hong et al. (2013) als Mitarbeitervariable betrachtet, die durch das Service-Klima beeinflusst wird. Ihr Modell spiegelt die Grundstruktur der Service-Profit Chain wider und sagt voraus, dass Service-Qualität zu Kundenzufriedenheit und Kundenzufrieden zum Erfolg der Organisation führt. Sie können ihr Modell metaanalytisch bestätigen und unterstreichen die Wichtigkeit des bereits in Kapitel 3.2.2 beschriebenen Konstrukts Service-Klima, das als wesentliches Bindeglied zwischen organisationsinternen und externen Service-Parametern identifiziert wurde.

Eine weitere Metaanalyse, die die Ergebnisse von 518 Primärstudien zusammenfasst, wurde von Hogreve et al. (2017) durchgeführt. Neben der Überprüfung der Grundstruktur der Service-Profit Chain untersuchen sie, welche Rolle Mitarbeiterzufriedenheit, Mitarbeiterbindung und die Zufriedenheit des Personals bei der Entstehung von Service-Qualität haben. Der Vergleich des Modells der Service-Profit Chain mit drei vorgeschlagenen Erweiterungen zeige, dass die Grundstruktur der Service-Profit Chain durch die Daten unterstützt wird, wobei der Modellfit für das reine Mediationsmodell gering ausfällt. Die getesteten Erweiterungen deuten darauf hin,

dass auch direkte Effekte zum Beispiel der Mitarbeiterzufriedenheit auf die Kundenzufriedenheit oder der Mitarbeiterbindung auf den Umsatz der Organisation beobachtet werden können.

3.4 Theoretisches Rahmenmodell und Hypothesen

In dieser Studie wurde die gewählte Operationalisierung von Service-Qualität (siehe Kapitel 3.3.4) mit den Stand der Forschung zu den Zusammenhängen zwischen Service-Qualität, Kundenzufriedenheit und dem Erfolg von Organisationen, die in Kapitel 3.3.5 beschrieben wurden, in ein Rahmenmodell integriert, das Abbildung 13 darstellt.

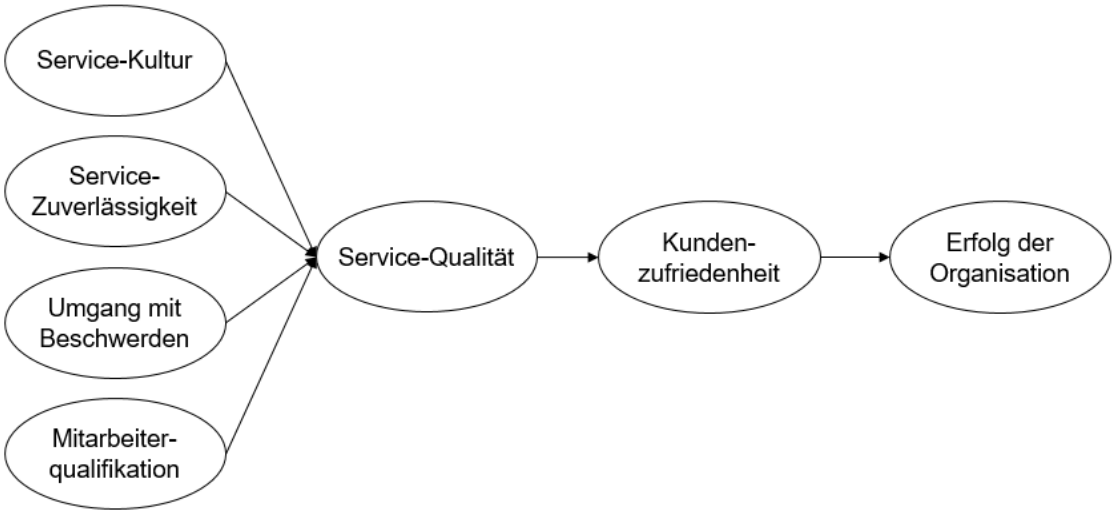


Abbildung 13 Rahmenmodell dieser Studie

Das Modell zeigt, dass die vier Komponenten Service-Kultur, Service-Zuverlässigkeit, Umgang mit Beschwerden und Mitarbeiterqualifikation gemeinsam als Indikatoren für Service-Qualität dienen. Inspiziert man das entwickelte Rahmenmodell genauer, so erkennt man, dass auch in diesem Modell, analog zur Grundstruktur der Service-Profit Chain, angenommen wird, dass der Zusammenhang zwischen Service-Qualität und dem Erfolg von Organisationen über die Kundenzufriedenheit vollständig mediiert wird (Gelade & Young, 2005; Heskett et al., 1994; Heskett et al., 1997; Heskett, Sasser & Wheeler, 2010; Hogreve et al., 2017; Hong et al., 2013).

4 Methoden

In diesem Kapitel wird zunächst die gewonnene Stichprobe beschrieben, die Operationalisierung der einzelnen Untersuchungsgegenstände dargestellt und anschließend geschildert, wie die einzelnen Instrumente innerhalb der Untersuchungsdurchführung zum Einsatz kamen. Abschließend werden die statistischen Analysemethoden beschrieben, die zur Datenanalyse eingesetzt wurden.

4.1 Beschreibung der Stichprobe

Um möglichst viele Experten zum Thema Service-Qualität zur Teilnahme an der Befragung zu motivieren, wurden in Kooperation mit dem TÜV Süd 6564 Personen aus deren Datenbank selektiert, die innerhalb ihrer Organisation direkt mit dem Thema Service-Qualität befasst sind. Von den ausgewählten Personen, die per Email zur Studie eingeladen wurden, nahmen N = 986 Personen an der Studie teil. 44,9 % der Befragten gaben an, Mitglied der Unternehmensleitung zu sein, 76,1 % übernahmen fachliche und 57,1 % disziplinarische Führungsverantwortung. Die meistvertretenen Branchen waren Maschinenbau (9,6 %), sonstige Industrie (9,6 %), Automotive (7,4 %), Lebensmittelindustrie (7,4 %) und Gesundheitswesen (7,1 %).

Da in der Studie sensible interne Unternehmensdetails erfragt wurden, war es wichtig, die vollständige Anonymität der Befragten zu garantieren. Aus diesem Grund und um die Befragung möglichst kurz zu halten, wurde darauf verzichtet, weitere soziodemografische Variablen wie zum Beispiel das Alter und Geschlecht der Befragten zu erfassen.

4.2 Operationalisierung der Untersuchungsgegenstände

In der psychologischen Forschung wird in der Regel mit latenten Konstrukten gearbeitet. Dahinter verbergen sich empirisch erkennbare Sachverhalte oder Eigenschaften, die nicht ohne weiteres direkt erfasst werden können. Zentrale Konstrukte in dieser Arbeit sind Service-Qualität, Kundenzufriedenheit und der Erfolg von Organisationen. In diesem Kapitel wird beschrieben, wie diese Konstrukte über die Entwicklung entsprechender Fragebogen-Items messbar gemacht wurden. Dieser Vorgang wird Operationalisierung genannt und erfordert eine sorgsame und gründliche Herangehensweise, da durch mangelhafte Operationalisierung die Aussagekraft und Generalisierbarkeit der Ergebnisse einer wissenschaftlichen Studie gefährdet wird.

4.2.1 Service-Qualität

Wie in Kapitel 3.3.3 und 3.3.4 beschrieben, existieren in der Fachliteratur zahlreiche Ansätze, Service-Qualität zu definieren und empirisch messbar zu machen. An der Vielzahl verschiedener theoretischer Definitionen und Operationalisierungsansätze lässt sich erkennen, dass sich in den verschiedenen akademischen Disziplinen, die sich dem Thema widmen, noch kein einheitliches Verständnis zum diesem Konstrukt herausgebildet hat. Die Heterogenität der Modelle und entwickelten Befragungsinstrumente lässt sich darauf zurückführen, dass bei deren Entwicklung unterschiedliche Ziele verfolgt werden.

Ein Ziel dieser Arbeit ist es, ein deutschsprachiges Befragungsinstrument zur Erfassung von Service-Qualität zu erarbeiten, das die Bestandteile bestehender Modelle für Service-Qualität integriert und darüber hinaus die Kriterien zur Zertifizierung des TÜV Süd berücksichtigt (TÜV Süd Management Service GmbH, 2017). Der entwickelte Fragebogen sollte nicht nur für akademische Zwecke nutzbar sein, sondern auch im Service-Management einer Organisation einsetzbar sein. Er sollte Information über die Service-Qualität im Allgemeinen, aber auch in verschiedenen Teilbereichen liefern und dem Management auf der Ebene der Einzelitems konkrete Ansatzpunkte für Verbesserungspotenziale anbieten.

Da Kundinnen und Kunden mit großen Teilen der Prozesse, die die Service-Qualität einer Organisation ausmachen, gar nicht in Kontakt kommen und diese nicht ausreichend präzise einschätzen können (Shemwell & Yavas, 1999), wurden Mitarbeiterinnen und Mitarbeiter, die ausreichend Überblick über den Bereich Service in der eigenen Organisation haben, als Zielgruppe des Instruments gewählt. Um möglichst viele Organisationen anzusprechen, wurde bei der Formulierung der Items darauf geachtet, dass alle von der Zielgruppe eingeschätzt und beantwortet werden können und nicht etwa zu branchenspezifisch sind.

Komplexe Phänomene in Organisationen werden in der Organisationspsychologie häufig als Konstrukte höherer Ordnung konzipiert. Ein Beispiel für ein solches Konstrukt ist organisationale Gerechtigkeit, das die drei Subdimensionen distributive Gerechtigkeit, prozedurale Gerechtigkeit und interaktionale Gerechtigkeit umfasst (Colquitt, Conlon, Wesson, Porter & Ng, 2001; Colquitt, 2001; Cropanzano & Greenberg, 1997). Solche Konstrukte mit Faktoren zweiter Ordnung, die auf sogenannten Second order-Modellen aufbauen, bieten in der praktischen Anwendung die Möglichkeit, mehrere Abstraktionsebenen zu unterscheiden (Cheung, 2008).

Wie in Abbildung 14 b dargestellt, wird Service-Qualität in dieser Studie ebenfalls als Konstrukt höherer Ordnung aufgefasst, das die Subdimensionen Service-Kultur (Beispielitem: Mitarbeiter, die sehr guten Service leisten, werden in unserem Unternehmen regelmäßig ausgezeichnet) Service-Zuverlässigkeit (Beispielitem: Wir halten mindestens in 90 Prozent aller Fälle unsere Servicestandards ein), Umgang mit Beschwerden (Beispielitem: Wir werten Beschwerden/Reklamationen für Verbesserungsmaßnahmen aus) und Mitarbeiterqualifikation

(Beispielitem: Die Weiterentwicklung unserer Mitarbeiter im direkten Kundenkontakt umfasst in hohem Umfang auch Elemente der sozialen Kompetenz) umfasst.

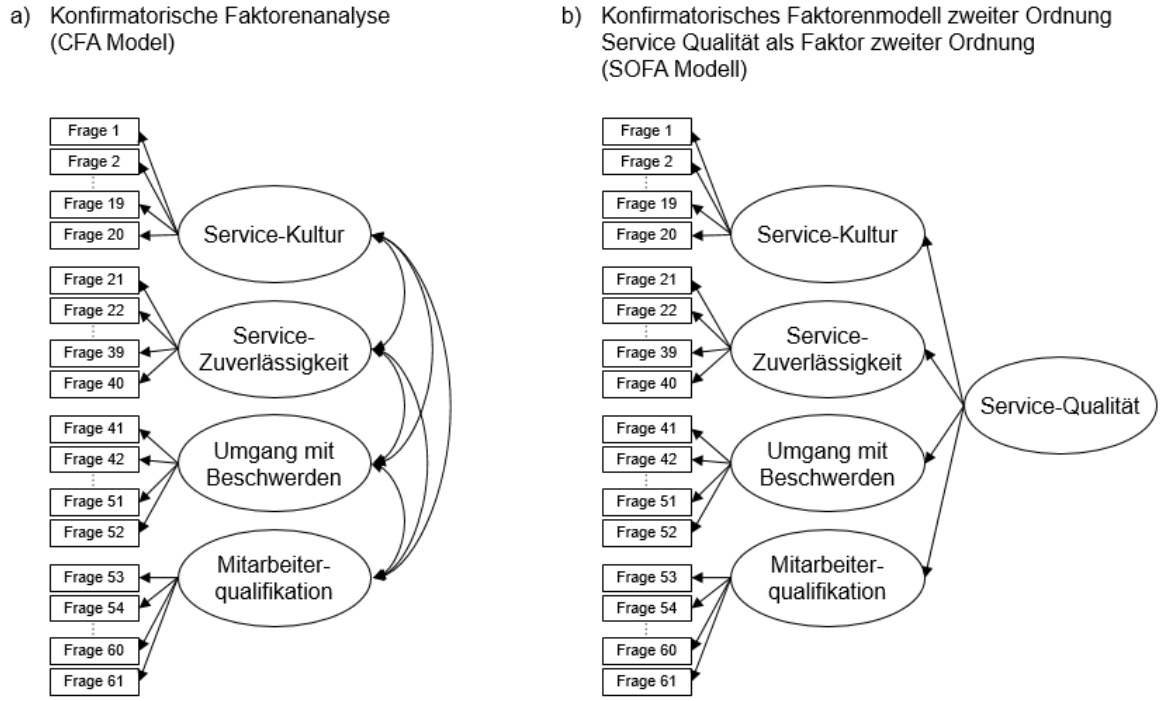


Abbildung 14 Modelle der konfirmatorischen Faktorenanalyse für Service-Qualität

Das allgemeine Gesamtniveau der Service-Qualität liefert Hinweise auf potenzielle Verbesserungspotenziale bei der Service-Qualität, die Ebene der Subfacetten kann aufzeigen, in welchem Bereich sich die Ansatzpunkte zur Verbesserung befinden, und auf der Ebene der Einzelitems kann exakt geprüft werden, bei welchem Aspekt von Service-Qualität sich die größten Verbesserungspotenziale zeigen.

In dieser Studie wurden die Kernelemente der bisherigen Forschung aufgegriffen und in qualitativen Interviews mit Fachleuten zum Thema Service diskutiert. In Kooperation mit dem TÜV Süd, einem Unternehmensberater und einem Professor für Arbeits-, Betriebs- und Organisationspsychologie wurden erste Aussagen zum Thema Service gesammelt. Die Auswertung dieser qualitativen Interviews zeigt, dass sich die zusammengetragenen Aussagen zu Serviceereignissen den vier Bereichen „Service-Kultur", „Service-Zuverlässigkeit", „Umgang mit Beschwerden" und „Mitarbeiterqualifikation" zuordnen lassen. Über mehrere Rückmeldungsschleifen wurden die Aussagen in Fragebogen-Items überführt, die anschließend auf Verständlichkeit und Eindeutigkeit überprüft wurden. Durch Expertenratings wurden anschließend 63 Fragebogen-Items bestimmt, die die Grundlage der entwickelten Skala zur Erfassung von

Service-Qualität bilden. Es wurde ein vierstufiges Antwortformat gewählt, bei dem nur die Extreme verbal mit „trifft nicht zu“ und „trifft voll zu“ verankert wurden. In einem Vortest der Skala als Papierfragebogen wiesen die Items keine ungewöhnlichen Verteilungseigenschaften und Item-Parameter auf. Eine vollständige Liste der Entwickelten Items befindet sich in Anhang A.

Die Reliabilität der Gesamtskala und der vier Subskalen, die in Tabelle 2 dargestellt wird, fiel gemessen an Cronbach's α (.85 > α <.96) und McDonald's ω (.84 > ω < .96) hoch aus.

Tabelle 2 Reliabilität der Gesamtsala und der Subskalen

Skala	α	ω
Service-Qualität Gesamt	.96	.96
Service Kultur	.91	.93
Service Zuverlässigkeit	.90	.92
Umgang mit Beschwerden	.85	.87
Mitarbeiterqualifikation	.84	.86

α = Cronbach's α; ω = McDonald's ω

4.2.2 Kundenzufriedenheit

Ein Überblick verschiedener Ansätze, Kundenzufriedenheit zu messen, wurde in Kapitel 3.2.5 gegeben. Um in dieser Studie die Fragebogenlänge und damit die Bearbeitungsdauer möglichst gering zu halten und so eine möglichst hohe Akzeptanz bei der befragten Zielgruppe zu erzielen, wurde nach einer Lösung gesucht, die die Kundenzufriedenheit möglichst ökonomisch erfasst. Zu Operationalisierungen von Kundenzufriedenheit wurde direkt nach den zentralen Elementen von Kundenzufriedenheit, der Wiederkaufbereitschaft, der Weiterempfehlungsbereitschaft und der Stabilität der Kundenbeziehung gefragt. Diese drei Items wurden auf einer fünfstufigen Skala, deren Extreme verbal mit „-2 viel schlechter“ und „+2 viel besser“ verankert waren, in Relation zu durchschnittlichen Organisationen aus dem gleichen Bereich eingeschätzt. Die Reliabilität dieser Skala für Kundenzufriedenheit fiel hoch aus (α = .84; ω = .84).

4.2.3 Erfolg der Organisationen

Um den Erfolg von Organisationen zu erfassen, können zahlreiche Indikatoren in Betracht gezogen werden. Häufig werden dazu die Ziele der Organisation und deren Erreichung betrachtet. Diese individuelle Betrachtung des Erfolgs von Organisationen ist sinnvoll, da zum Beispiel die wirtschaftliche Situation in verschiedenen Branchen und Geschäftsbereichen unterschiedliche Herausforderungen und Erwartungen mit sich bringt. Eine solche auf den Einzelfall ab-

gestimmte Operationalisierung des Erfolgs einer Organisation ist sehr aufwendig und macht den Vergleich des Erfolgs über verschiedene Organisationen hinweg schwierig. Da in dieser Studie Organisationen aus sehr heterogenen Bereichen adressiert wurden, wurde entschieden, als Indikatoren für den Erfolg einer Organisation deren Umsatz und Gewinn zu nutzen. Da Umsatz und Gewinn von zahlreichen organisations- und branchenspezifischen Faktoren abhängen, wurde eine Einschätzung im Vergleich zu durchschnittlichen Unternehmen in der gleichen Branche auf einer fünfstufigen Skala, deren Pole mit „-2 viel schlechter“ und „+2 viel besser“ verankert wurden, erfasst. Fasst man diese beiden Items zu einer Skala zusammen, die den Erfolg einer Organisation abbildet, so fällt die Reliabilität dieser Skala zufriedenstellend aus ($\alpha = .76$; $\omega = .76$).

4.3 Durchführungsbeschreibung

Zur Erfassung der Daten dieser Studie wurde ein Onlinebefragungssystem entwickelt, das im Wesentlichen auf den Internettechnologien HTML, PHP und MySQL basiert. Die Entwicklung dieses Systems war notwendig, weil die etablierten Onlinebefragungssysteme nicht in der Lage waren, den gewünschten Funktionsumfang abzudecken. Insbesondere die Anforderungen, dass Nutzer und Nutzerinnen sich mehrfach einloggen können und dass die Daten von mehreren Messzeitpunkten im System abgelegt und direkt und automatisiert weiterverarbeitet werden, um daraus individuelle, dynamische Rückmeldungsgrafiken zu erzeugen, machte eine Eigenentwicklung erforderlich. Das entwickelte Onlinebefragungssystem zeigt dem Nutzer über einen Fortschrittsbalken den Bearbeitungsstand der Umfrage an und ermöglicht es den Teilnehmenden, die Befragung zu unterbrechen und zu einem späteren Zeitpunkt fortzusetzen.

In das entwickelte System wurden die E-Mail-Adressen der Stichprobe eingefügt, zudem wurde zu jeder Emailadresse ein Passwort generiert, um sicherzustellen, dass sich nur Personen aus der gewählten Stichprobe an der Studie beteiligen konnten. Das System ermöglicht den automatisierten Versand von personalisierten Emails an die Versuchspersonen, was zum Einladen und Erinnern der Personen aus der Stichprobe wurde. Die ausgewählten Personen erhielten eine E-Mail mit ihren Zugangsdaten und einem personalisierten Direktlink, über den sie sich ohne zusätzliche Eingabe ihrer Benutzerkenndaten im Befragungssystem anmelden konnten. Nach den einleitenden Instruktionen wurden den Versuchspersonen sequenziell alle Items der Skala zur Erfassung von Service-Qualität dargeboten. Im Anschluss wurden Kundenzufriedenheit, Erfolg der Organisation und noch einige Informationen über die Unternehmensgröße und die Branche erhoben. Abschließend erhielten die Teilnehmenden eine Auswertung ihres Fragebogens in Form eines Liniendiagramms, das sie ausdrucken oder als PDF-Dokument abspeichern konnten.

Im Vorfeld der Hauptuntersuchung wurden in einer Pilotstudie die für die Hauptuntersuchung programmierten Funktionen und Internetseiten vorgetestet. Es wurde überprüft, ob der Versand der Einladungs- bzw. Erinnerungsemails funktioniert, die erstellten Webseiten technisch einwandfrei funktionieren, auf verschiedenen Endgeräten gut les- und nutzbar dargestellt werden und die Instruktionen verständlich sind, so dass eine reibungslose Bearbeitung ohne weitere Unterstützung gewährleistet ist. Die Erfahrungen, die im Vortest gesammelt wurden, führten zu kleinen Optimierungen der Instruktionstexte und deuteten insgesamt darauf hin, dass die Datenerhebung der Hauptuntersuchung über das entwickelte System durchgeführt werden kann. Diese Voruntersuchung ermöglichte zusätzlich, anhand erster Daten abzuschätzen, ob die gewählten Fragebogenitems im Hinblick auf ihre psychometrischen Eigenschaften geeignet sind, um die gewünschten Konstrukte zu erfassen. Die Daten der Hauptuntersuchung dieser Studie wurden mit dem entwickelten System erfasst und für die weitere Datenaufbereitung und -analyse gespeichert und exportiert.

4.4 Statistische Methoden

Zur Prüfung der Hypothesen dieser Arbeit wurden die erfassten Daten mit der Open Source-Statistik-Software R in der Version 3.6.1 verarbeitet und analysiert (*R,* 2010). Die eingebunden Datenvisualisierungen wurden, wenn nicht anders angemerkt, mit dem R-Paket „ggplot2“ (Wickham, 2016) generiert. Sowohl die konfirmatorischen Faktorenanalysen als auch das Strukturgleichungsmodelle wurden mit dem R-Paket „lavaan“ in der Version 0.6-5 spezifiziert und geschätzt (Rosseel, 2012). Bei beiden Analysen wurde davon ausgegangen, dass es durch das gewählte Antwortformat, bei dem lediglich die Extrempole mit „trifft nicht zu“ und „trifft voll zu“ verbal verankert waren, nicht gelungen ist, ein Intervallskalenniveau bei den Einzelitems zu erreichen (Sellbom & Tellegen, 2019). Dies würde nur dann gelten, wenn die Abstände zwischen den Antwortalternativen als gleich groß zu betrachten wären. Sowohl theoretisch als auch nach Betrachtung der empirischen Verteilung der Antwortdaten wurde ein ordinales Skalenniveau angenommen. Bei der Spezifikation der Modelle wurden die Indikator-Items deshalb als ordinale Items betrachtet und als Parameterschätzer WLSMV bzw. DWLS genutzt (Beauducel & Herzberg, 2006; DiStefano & Morgan, 2014; Heflin, Sandberg & Rafail, 2009; Hutchinson & Olmos, 1998; Li, 2016a, 2016b; Savalei, 2014; Sellbom & Tellegen, 2019). Für weitere Analysen und Visualisierungen wurden spezialisierte R-Pakete genutzt, auf die an der jeweiligen Stelle im Text verwiesen wird.

4.4.1 Bestimmung der Reliabilität

Um die Reliabilität der Skalen Service-Qualität zu beurteilen, wurde – im Blick auf Vergleichbarkeit mit anderen Studien – Cronbach's α berichtet (Cronbach, 1951). Wie methodische Studien gezeigt haben, ist die Verwendung von Cronbachs's α als Schätzer für die Reliabilität problematisch, weil die Annahmen häufig verletzt sind (Dunn, Baguley & Brunsden, 2014; McNeish, 2017). Aus diesem Grund wurde als weiterer Indikator für die Reliabilität der Skalen McDonald's ω mit dem R-Paket „psych" berechnet (Revelle, 2010; Revelle & Condon, 2019).

4.4.2 Modelle der konfirmatorischen Faktorenanalyse

Um zu prüfen, ob die theoretisch angenommene faktorielle Struktur der Items der Skala Service-Qualität mit ihren vier Subfacetten, die in Abbildung 14 dargestellt wurde, anhand der empirischen Daten untermauert werden kann, wurde zunächst eine konfirmatorische Faktorenanalyse erster Ordnung (CFA) durchgeführt. Im CFA-Modell, das in Abbildung 14 a dargestellt wurde, sind die Items gemäß der theoretischen Annahmen den vier Faktoren erster Ordnung: Service-Kultur, Service-Zuverlässigkeit, Umgang mit Beschwerden und Mitarbeiterqualifikation zugeordnet. Die Korrelation diese Faktoren erster Ordnung wurde frei geschätzt und berichtet. Dieses CFA-Modell erster Ordnung wurde verglichen mit dem in Abbildung 14 b dargestellten Modell, das einem Faktor zweiter Ordnung enthält (SOFA), der als Service-Qualität interpretiert wird. Dieser Faktor zweiter Ordnung soll die Zusammenhänge der Faktoren erster Ordnung erklären, die in diesem Modell nicht mehr miteinander korreliert sind. Um die Modellgüte dieser beiden Modelle besser einschätzen zu können, wurde abschließend ein CFA-Modell geschätzt und berichtet, in dem alle Items der Skala Service-Qualität auf nur einem latenten Faktor laden, der als Gesamt-Service-Qualität interpretiert werden kann. Dieses letzte Modell blendet die vier Subdimensionen aus, die bei der Konstruktion der Skala zur Erfassung von Service-Qualität grundlegend waren. In allen Modellen der konfirmatorischen Faktorenanalyse wurde zur Skalierung der latenten Variable und zur Schätzung der Faktorladungen für alle Items die Varianz der latenten Dimensionen auf 1 festgelegt (MacCallum, 1995; McDonald & Ho, 2002; Thompson, 2004).

4.4.3 Strukturgleichungsmodelle zur Prüfung des Rahmenmodells

Zur Prüfung der Hypothesen dieser Arbeit, die im Rahmenmodell zusammengefasst dargestellt sind (s. Kap. 3.4, Abb. 13), wurden ein Strukturgleichungsmodell genutzt. Der linke Teil des Rahmenmodells entspricht dem CFA-Modell mit einem Faktor zweiter Ordnung (SOFA). Analog zu den Modellen der CFA wurde die Varianz der latenten Dimensionen auf 1 festgelegt (MacCallum, 1995; McDonald & Ho, 2002; Thompson, 2004). Die latente Variable Kunden-

zufriedenheit wurde über drei Indikatorvariablen spezifiziert, wie in Kapitel 4.2.2 beschrieben. Um den Erfolg der Organisation im Strukturgleichungsmodell abzubilden, wurden die zwei in Kapitel 4.2.3 beschriebenen Indikatoren genutzt.

4.4.4 Beurteilung der Modellgüte

Um die Güte eines Strukturgleichungsmodels zu beurteilen, werden in der Literatur verschiedene Indikatoren empfohlen (Schreiber, Nora, Stage, Barlow & King, 2006). Basierend auf der Abweichung zwischen der empirischen Varianz-Kovarianz-Matrix und der vom Strukturgleichungsmodell implizierten Varianz-Kovarianz-Matrix kann der χ^2-Wert und die zugehörige Irrtumswahrscheinlichkeit als p-Wert bestimmt werden. Stellt dieser Parameter keine signifikanten Abweichungen zwischen den beiden Matrizen fest, spricht dies dafür, dass das Strukturgleichungsmodell mit seinen geschätzten Parametern die empirischen Interkorrelationen zwischen den Items gut nachbilden kann, und es kann von einer guten Modellgüte ausgegangen werden. Bei größeren Stichproben steigt die statistische Power dieses Tests an, so dass auch kleinste, möglicherweise irrelevante Abweichungen signifikant werden. Um diese Tatsache zu berücksichtigen, wurde der Quotient aus dem χ^2-Wert und den damit assoziierten Freiheitsgraden als weiteres Maß für die Beurteilung der Modellgüte herangezogen. Je kleiner dieser Quotient ausfällt, umso besser ist der Modellfit einzuschätzen. Es existieren keine absoluten Standards zur Interpretation dieses Quotienten, ein Wert zwischen 2 und 3 gilt als Indikator für einen guten oder akzeptablen Modellfit (Schermelleh-Engel, Moosbrugger & Müller, 2003). Der χ^2-Test berücksichtigt nur die Abweichungen zwischen der empirischen Varianz-Kovarianz-Matrix und der modellimplizierten Varianz-Kovarianz-Matrix, weshalb komplexere Modelle mit mehreren zu schätzenden Parametern, die die empirische Varianz-Kovarianz-Matrix besser nachbilden können, von diesem Test bevorzugt werden. Wird beim Beurteilen der Modellgüte ausschließlich auf diesen Test vertraut, verstößt dies gegen das Prinzip der Parsimonität bzw. Sparsamkeit (Epstein, 1984). Als Alternativen werden deskriptive Maße der inkrementellen Anpassungsgüte, wie der Tucker-Lewis Index(TLI) und der Comparative Fit Index (CFI), berichtet, die komplexere Modelle bestrafen. Von einem guten Modellfit kann ausgegangen werden, wenn der TLI einen Wert > 0,9 und der CFI einen Wert > 0,95 aufweist (Hu & Bentler, 1998, 1999; Marsh, Hau & Wen, 2004). Ein weiteres Maß zur Beurteilung der Modellgüte, das auch in dieser Arbeit berichtet wird, ist der Root Mean Square Error of Approximation (RMSEA). Für dieses Maß, das die Diskrepanz bezogen auf die Approximation schätzt, gelten Werte < 0,05 als Indikator für einen guten Modellfit, Werte > 0,05 und < 0,08 als Hinweis auf adäquate bzw. mäßige Modellpassung und größere Werte als Zeichen für schlechte Modellpassung (Hu & Bentler, 1999; Kaplan, 2000). Neben dem Punktschätzer für den RMSEA werden in dieser Arbeit auch Intervallgrenzen eines 90 % Konfidenzintervalls für diesen Parameter berichtet.

Ein weiteres absolutes Gütemaß, das berichtet wird, ist der Standardized Root Mean Square Residual (SRMR). Dieser sollte möglichst klein ausfallen, wobei Werte < 0,05 als gute und Werte < 0,10 als akzeptable Modellpassung interpretiert werden können (Barrett, 2007; Jackson, Gillaspy & Purc-Stephenson, 2009; Schermelleh-Engel et al., 2003).

5 Ergebnisse

In diesem Teil der Arbeit werden zunächst die deskriptiven Statistiken der Befragungsdaten dargestellt. Anschließend wird anhand von Zusammenhangsmaßen die dimensionale Struktur auf der Ebene der Items näher untersucht. Eine konfirmatorische Faktorenanalyse untersucht, wie gut es gelungen ist, die theoretisch abgeleiteten Facetten des Konstrukts Service-Qualität empirisch abzubilden. Abschließend wird das theoretische Rahmenmodell, das in Kapitel 3.4 entwickelt wurde, mithilfe eines linearen Strukturgleichungsmodells geprüft.

5.1 Analyse der Fragebogendaten

In diesem Abschnitt werden die Antwortdaten aus der Onlinebefragung analysiert. Dazu wird zunächst dargestellt, wie häufig die Versuchspersonen die einzelnen Antwortkategorien der Fragebogenitems gewählt haben. Anschließend wird die Verteilung der Antworten im Rahmen einer Item-Analyse untersucht und anhand deskriptiver Statistiken abgebildet. Danach werden die Item-Interkorrelationen berechnet und die faktorielle Struktur der Daten dargestellt und diskutiert.

5.1.1 Häufigkeitsverteilung der Antworten

Um einen Eindruck der Häufigkeitsverteilung der Antwortdaten zu den Items der Skala Service-Qualität zu bekommen, wird in Abbildung 15 für jedes Item ein Balkendiagramm dargestellt.

Wie auf den ersten Blick deutlich wird, gibt es Items, bei denen alle Antwortkategorien einigermaßen gleich häufig ausgewählt wurden (z. B. Frage 50), bei anderen Items wurde die Antwortkategorie „4 = trifft voll zu“ eindeutig am häufigsten gewählt (z. B. Frage 48). Die Antwortkategorie „1 = trifft nicht zu“ wurde bei einigen Items sehr selten genutzt. Die Antwortkategorie, die in Abbildung 15 mit „NA“ bezeichnet wird, steht für fehlende Werte. Betrachtete man diese Kategorie über die Einzel-Items hinweg, sticht kein Item heraus, das besonders häufig nicht beantwortet wurde. Weil nicht alle Teilnehmenden den Fragebogen vollständig bis zum Ende bearbeitet haben, kann beobachtet werden, dass Items, die in der linear dargebotenen Befragung zu einem späteren Zeitpunkt präsentiert wurden, einen höheren Anteil an fehlenden Werten aufwiesen.

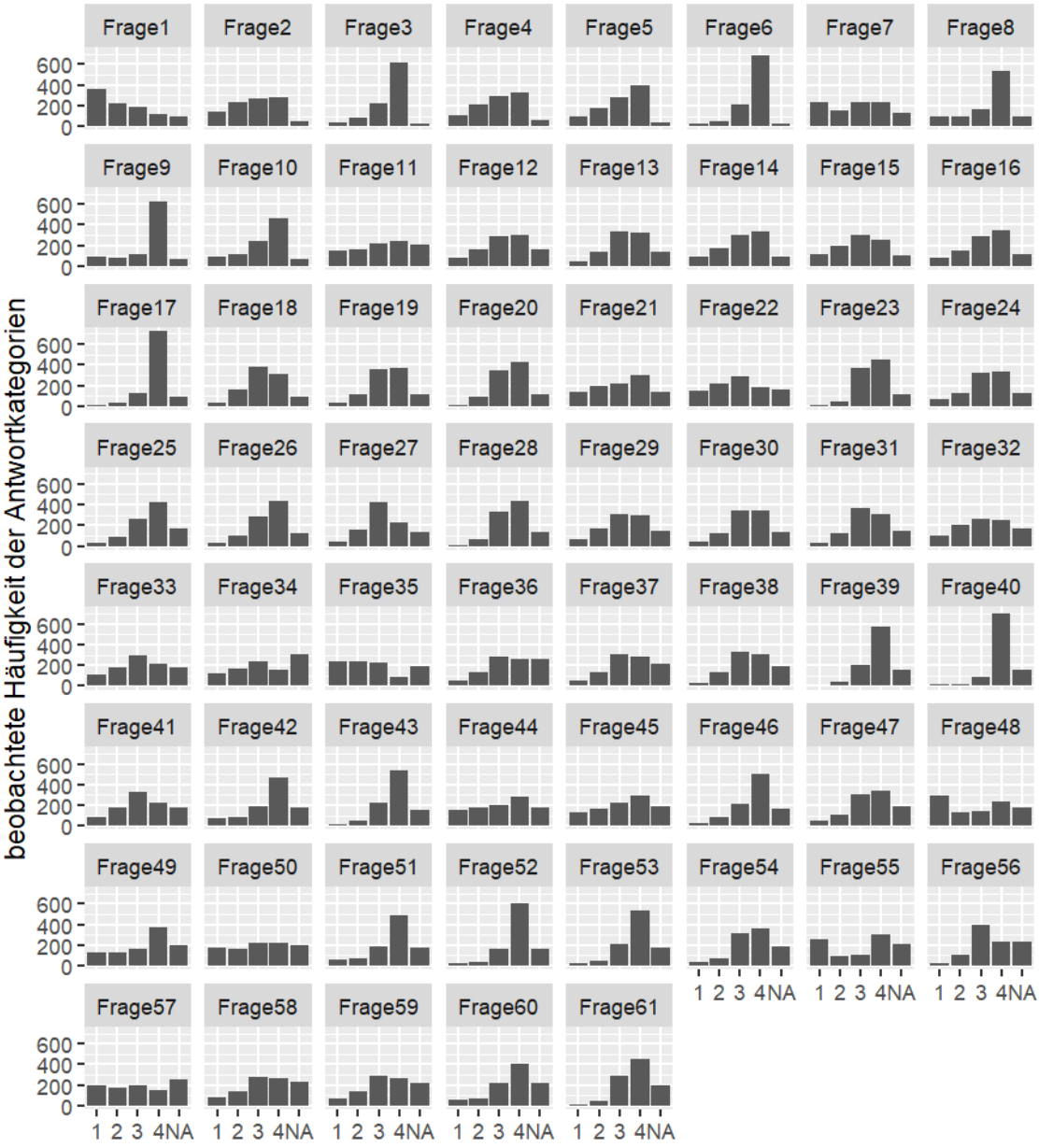


Abbildung 15 Häufigkeitsverteilung der Antworten auf die Items der Skala Service-Qualität, durch das R-Paket „ggplot2“ dargestellt (1 = trifft nicht zu; 4 =trifft voll zu; NA = not available / fehlende Werte)

Die Häufigkeitsverteilung der Antwortdaten zu den Indikator-Items für den Erfolg der Organisation und der Kundenzufriedenheit werden in Abbildung 16 dargestellt.

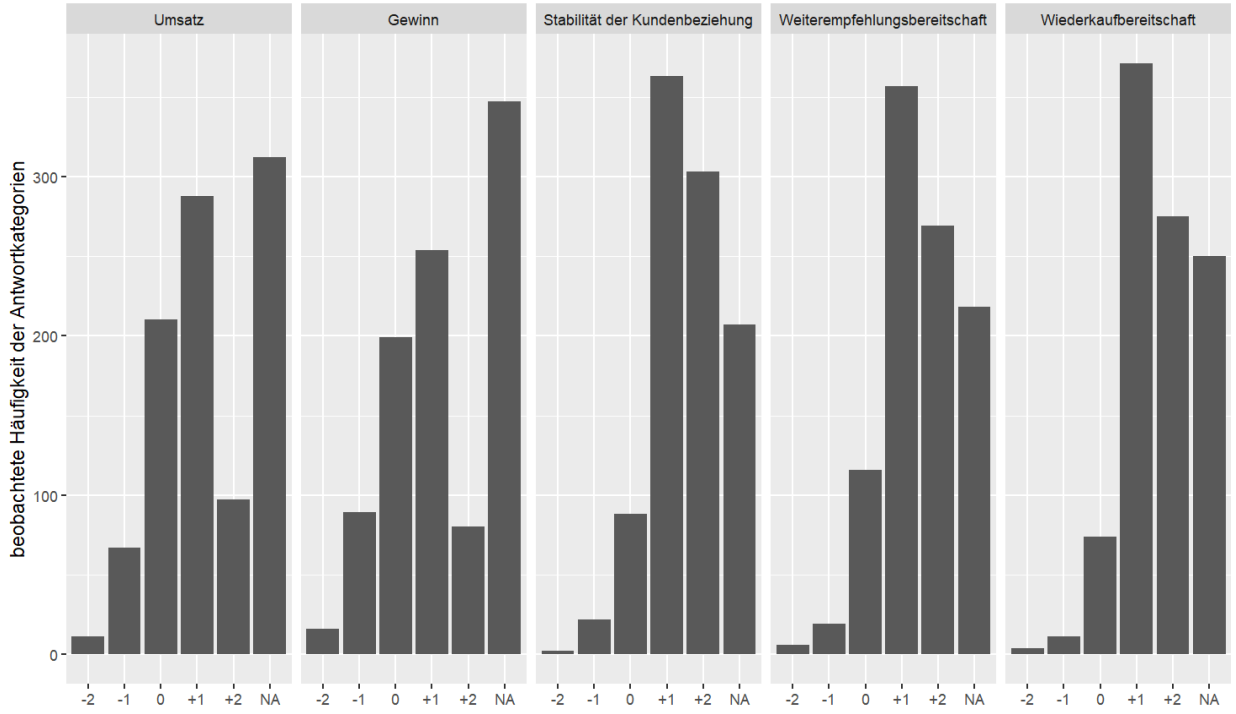


Abbildung 16 Häufigkeitsverteilung der Indikator-Items für Erfolg und Kundenbeziehung, die auf einer fünfstufigen Skala, deren Extreme verbal mit „-2 viel schlechter“ und „+2 viel besser“ verankert waren, erfasst wurden; NA = fehlender Wert

Bei den Indikatoren für Kundenzufriedenheit (Stabilität der Kundenbeziehung, Weiterempfehlungsbereitschaft und Wiederkaufbereitschaft) wurde die Antwortkategorie „+2 viel besser“ häufiger gewählt als bei den Indikatoren für den Erfolg der Organisation (Gewinn und Umsatz). Bei den Indikatoren für den Erfolg der Organisation wurden mehr fehlende Werte (NA) beobachtet als bei den Indikatoren für Kundenzufriedenheit.

5.1.2 Maße der zentralen Tendenz und der Streuung

Um einen Überblick über das Antwortverhalten und die psychometrischen Eigenschaften der Items der Skala Service-Qualität zu erhalten, gibt Tabelle 9 im Anhang A eine Übersicht über die Item-Mittelwerte (*M*), die als Schwierigkeitsindikator interpretiert werden können, die Standardabweichung (*SD*), als ein Maß für die Variabilität im Antwortverhalten, und die Trennschärfe ($r_{i(t-i)}$), ermittelt als Korrelation eines Einzelitems mit dem Testwert, der sich ergibt, wenn alle anderen Items der Skala zusammengefasst werden. Tabelle 9 enthält darüber hinaus Informationen über die Item-Namen, die genaue Formulierung der Items im Fragebogen und die Anzahl der vorliegenden Antworten zu jedem Einzel-Item.

Um die Informationsfülle in Tabelle 9 leichter erfassbar zu machen, stellt Abbildung 17 die Verteilung der Mittelwerte der Items als Histogramm dar.

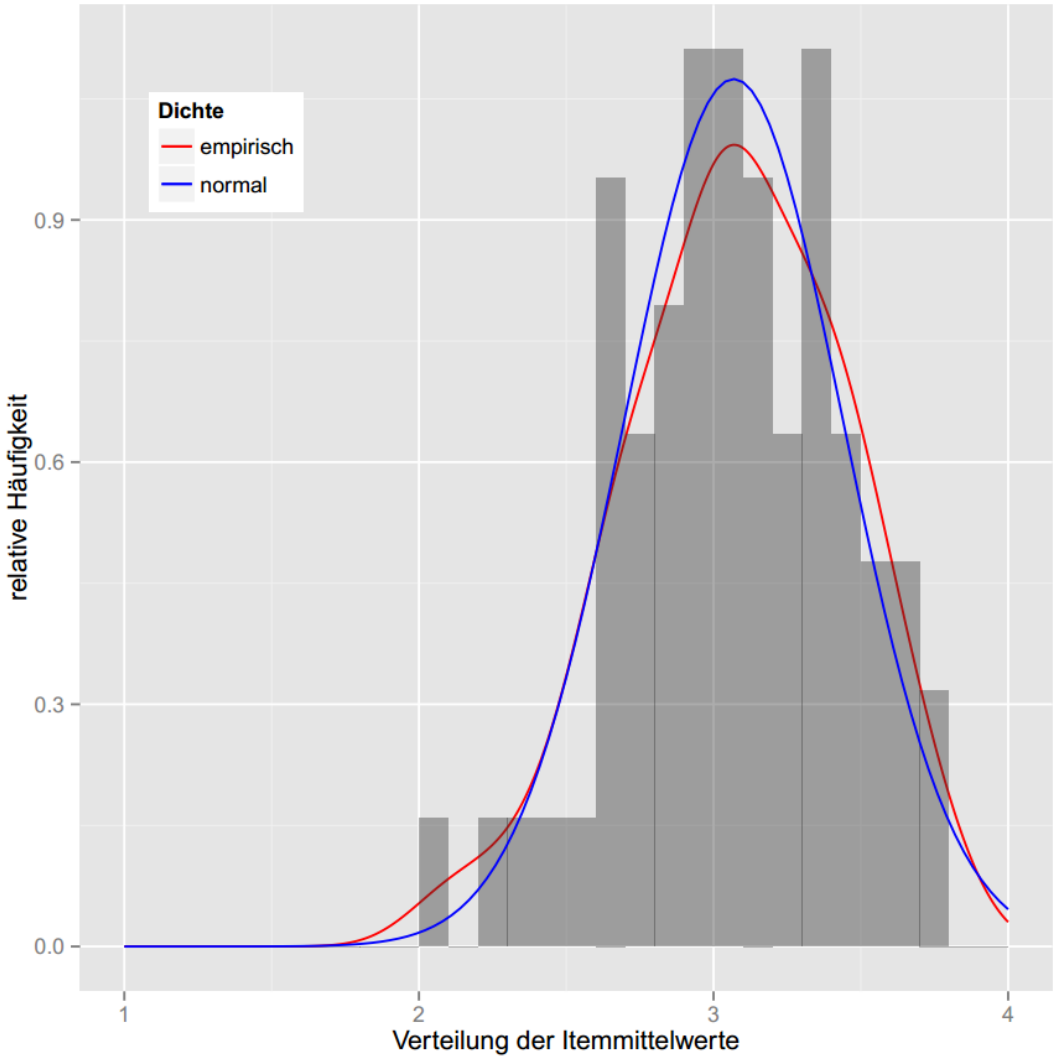


Abbildung 17 Häufigkeitsverteilung der Item-Mittelwerte. Die rote Kurve entspricht der empirischen Dichteverteilung. Die blaue Kurve bildet eine Normalverteilung mit dem Mittelwert und der Standardabweichung der empirischen Daten ab.

Abbildung 17 verdeutlicht, dass den Aussagen der einzelnen Fragebogen-Items überwiegend zugestimmt wurde. Der Mittelwert aller Item-Mittelwerte fällt mit $M = 3.07$ relativ hoch aus, woran ebenfalls erkennbar wird, dass von der vierstufigen Antwortskala bei den meisten Items hauptsächlich die höheren Kategorien ausgewählt wurden.

Den geringsten Mittelwert weist Item 1 ($M = 2.08$; $SD = 1.07$) mit dem Wortlaut: „Mitarbeiter, die sehr guten Service leisten, werden in unserem Unternehmen regelmäßig ausgezeichnet“

auf. Den höchsten Mittelwert (M = 3,78; SD = 0.58) hat Item 40 mit dem Wortlaut: „Unseren Kunden entstehen keine erheblichen zusätzlichen Kosten für die Kontaktaufnahme mit uns“. Die Tatsache, dass die Verteilung der Item-Mittelwerte nahezu perfekt einer Normalverteilung entspricht, was an der geringen Abweichung zwischen der roten und der blauen Kurve in Abbildung 17 abgelesen werden kann, korrespondiert mit den Aussagen des zentralen Grenzwertsatzes (Lindeberg, 1922).

Die Verteilung der Standardabweichungen der Items, die in Abbildung 18 grafisch dargestellt wird, verdeutlicht, dass auch diese annähernd normalverteilt sind.

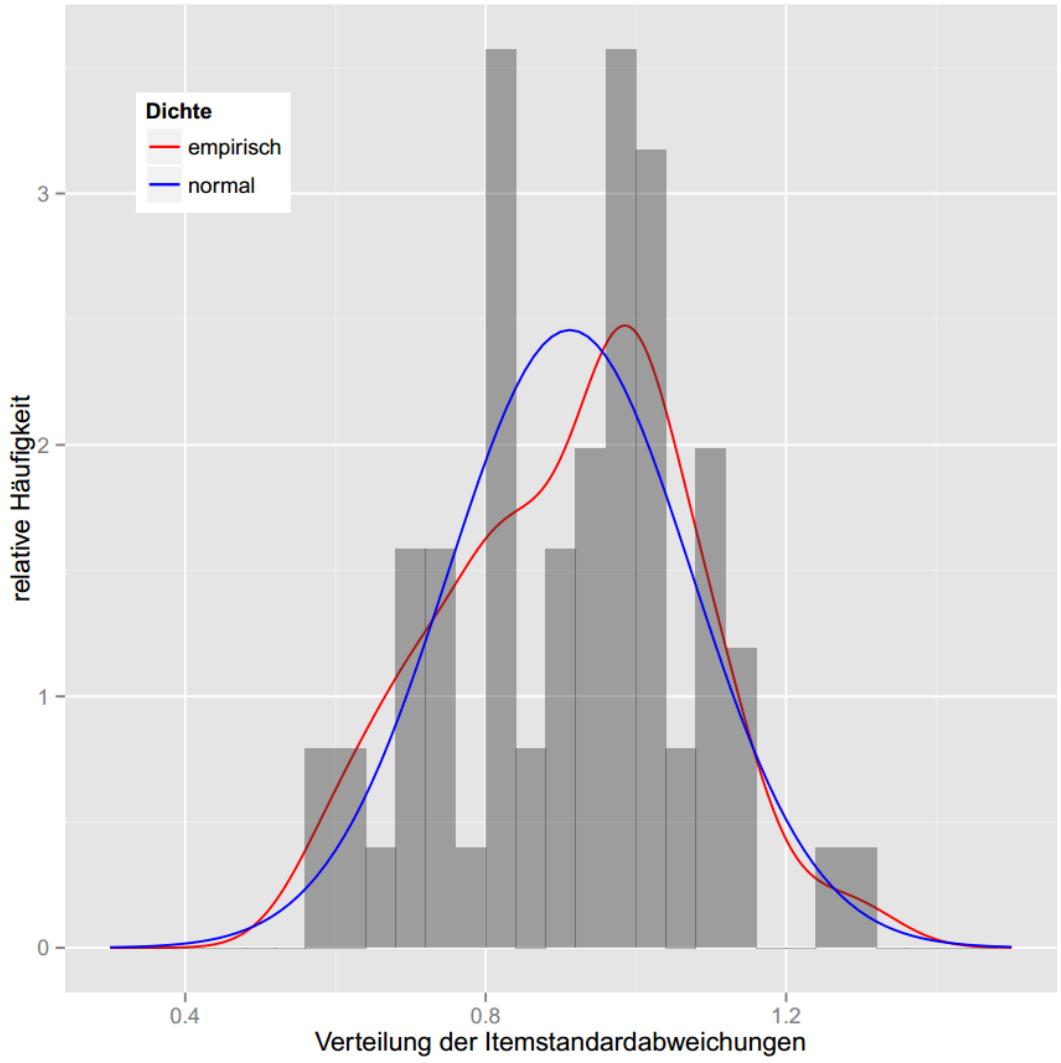


Abbildung 18 Häufigkeitsverteilung der Standardabweichungen. Die rote Kurve entspricht der empirischen Dichteverteilung. Die blaue Kurve bildet eine Normalverteilung mit dem Mittelwert und der Standardabweichung der empirischen Daten ab.

Einige Items stechen jedoch mit besonders hohen bzw. besonders niedrigen Standardabweichungen hervor. Zum Beispiel fällt Item 55 mit der höchsten beobachteten Standardabweichung von *SD* = 1.30 auf. Der Wortlaut dieses Items ist: „Wir führen mindestens alle zwei Jahre eine Mitarbeiterzufriedenheitsbefragung durch". Item 17 mit der Formulierung: „Kundenzufriedenheit ist ein wesentlicher Bestandteil unseres Unternehmensleitbildes" weist die geringste Standardabweichung (*SD* = 0.58) auf.

Die Indikatoren für den Erfolg der Organisation (Umsatz und Gewinn) und die Indikatoren für die Kundenzufriedenheit (Stabilität der Kundenbeziehung, Weiterempfehlungsbereitschaft und Wiederkaufbereitschaft) wurden bei der Befragung auf einer fünfstufigen Skala, deren Extreme verbal mit „-2 viel schlechter" und „+2 viel besser" verbal verankert waren, erfasst. Für die Datenanalyse wurden die fünf Antwortkategorien mit den Werten 1 bis 5 kodiert. Tabelle 3 stellt die Mittelwerte und Standardabweichung dieser Indikatorvariablen dar.

Tabelle 3 Mittelwerte und Standardabweichungen der Indikatoren für den Erfolg der Organisation und der Kundenzufriedenheit

Konstrukt	Variable	*M*	*SD*	*n*
Erfolg	Umsatz	3.58	0.91	673
Erfolg	Gewinn	3.46	0.96	638
Kundenzufriedenheit	Stabilität der Kundenbeziehung	4.21	0.77	778
Kundenzufriedenheit	Weiterempfehlungsbereitschaft	4.13	0.81	767
Kundenzufriedenheit	Wiederkaufbereitschaft	4.23	0.73	735

M = Mittelwert; *SD* = Standardabweichung; *n* = Stichprobenumfang

Die Indikatoren für Kundenzufriedenheit weisen höhere Mittelwerte und leicht geringere Standardabweichungen auf als die Indikatoren für den Erfolg der Organisation.

5.1.3 Interkorrelationen der Items

In diesem Abschnitt wird untersucht, wie die Antworten auf die einzelnen erfassten Items untereinander korrelieren und ob die angenommene Faktorenstruktur sich in den erfassten Daten empirisch abbildet. Um die dimensionale Struktur, die den einzelnen Fragen zugrunde liegt, zu überprüfen, wurde zunächst ein Blick auf die Interkorrelationen der Einzel-Items geworfen. Tabelle 10 im Anhang B enthält die Korrelationsmatrix, in der die empirischen linearen Produkt-Moment-Korrelationen zwischen allen Items der Skala zur Erfassung von Service-Qualität dargestellt wurden. Zur Beurteilung der Interkorrelationen von 61 Fragen muss eine große

Anzahl von einzelne Korrelationskoeffizienten (betrachtet werden. In Tabelle 10 im Anhang ist die entsprechende Korrelationsmatrix abgebildet. Da diese Darstellung die Beurteilung der Struktur der Korrelationsmuster nur schwer ermöglicht, wird in Abbildung 19 auf eine farbkodierte Darstellung der Korrelationen zurückgegriffen, in der blaue Farbe auf einen hohen positiven, weiße Farbe auf keinen und rote Farbe auf einen negativen Zusammenhang hinweist.

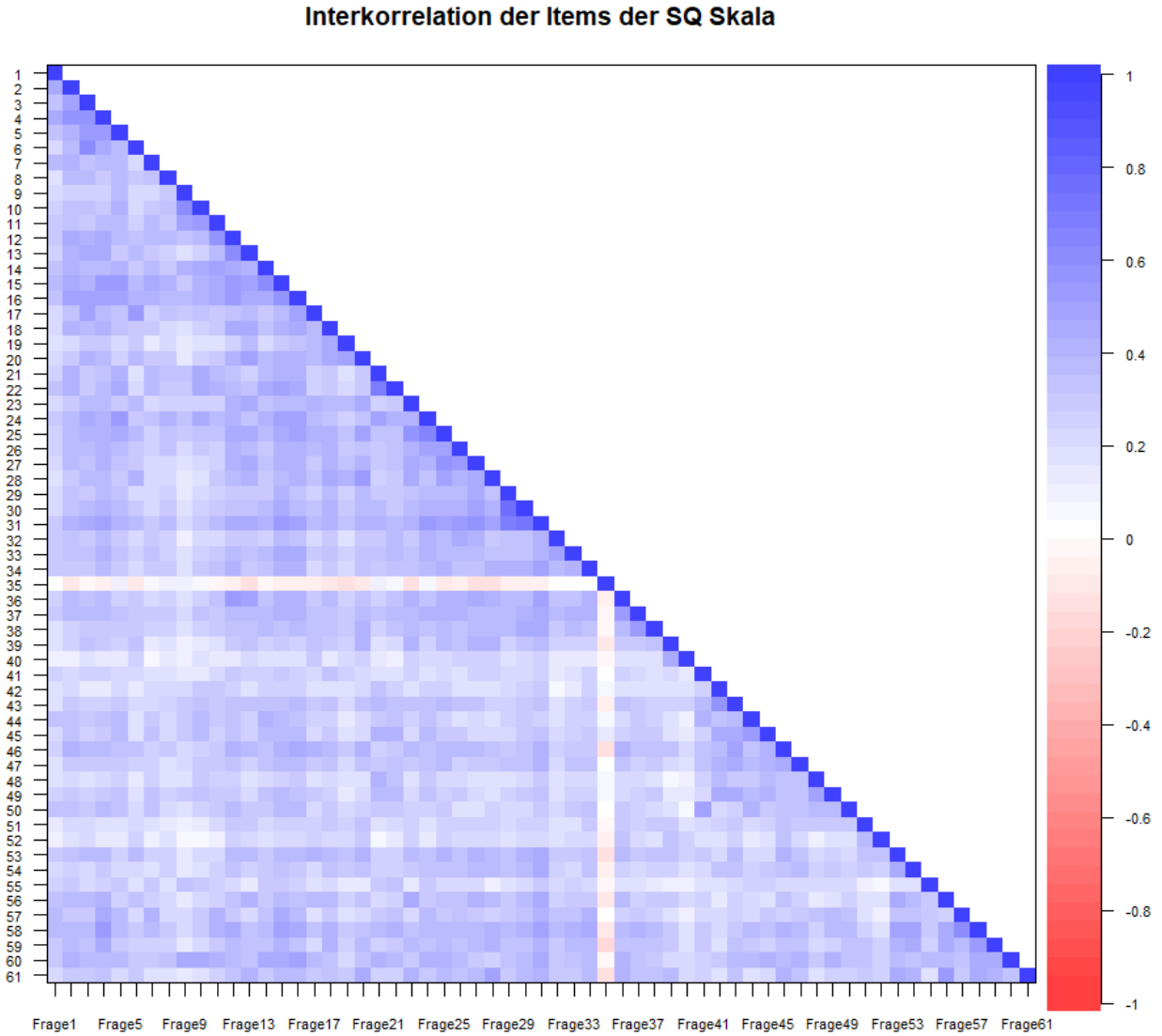


Abbildung 19 Interkorrelation der Einzelitems farbkodiert dargestellt durch die Funktion cor.plot() aus dem R-Paket „psych“ (Revelle, 2010); blau = hohe positive Korrelation, weiß = keine Korrelation, rot = hohe negative Korrelation

Auf den ersten Blick fällt, neben der Hauptdiagonalen, in der die perfekte Interkorrelation der Items mit sich selbst als blaue Linie dargestellt ist, Frage 35 auf, die als einziges Item negative Korrelationen zu den anderen Items aufweist. Abbildung 20, in der die beobachteten Korrelationen als Netzwerk visualisiert werden, verdeutlicht diese Sonderrolle von Frage 35. In dieser Abbildung werden einzelnen Items als Netzwerkknoten dargestellt. Die Zahl, die in jedem Netzwerkknoten abgebildet ist, steht für die Item-Nummer. Die Interkorrelationen werden in dieser Darstellungsform über Verbindungslinien zwischen den einzelnen Knotenpunkten dargestellt, wobei dicke grüne Linien für eine hohe positive und dicke rote Linien für eine hohe negative Korrelation stehen. Die Anordnung der Netzwerkknoten erfolgt hier nach dem Fruchterman-Reingold-Algorithmus (Fruchterman & Reingold, 1991).

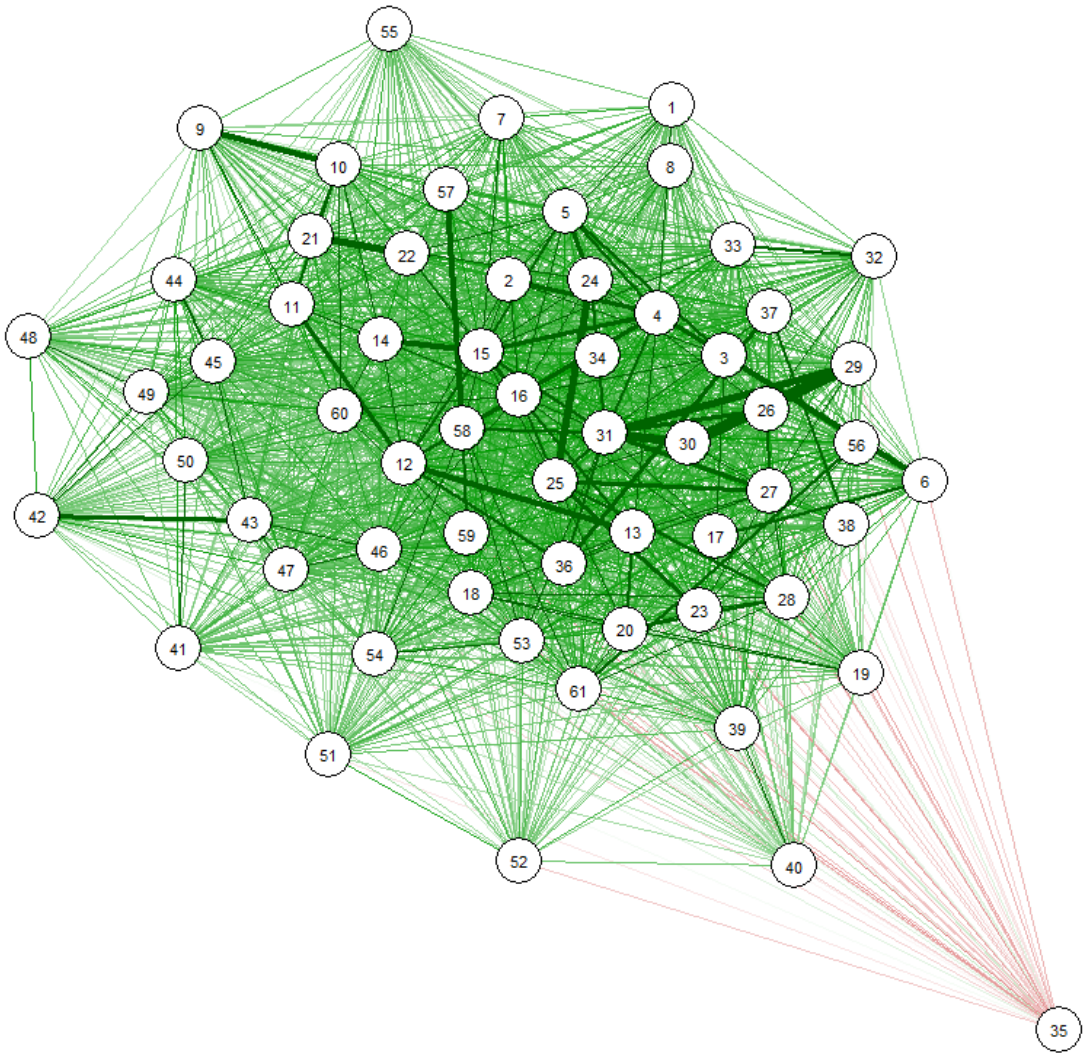


Abbildung 20 Visualisierung der Interkorrelationen der Items als Netzwerkdarstellung durch das R-Paket „Qgraph“ (Epskamp, Cramer, Waldorp, Schmittmann & Borsboom, 2012); Anordnung der Items nach dem Fruchterman-Reingold Algorithmus (Fruchterman & Reingold, 1991)

Um einen besseren Eindruck von der faktoriellen Substruktur der Items zu erhalten, kann bei einer netzwerkartigen Darstellung der Item-Interkorrelationen die Anordnung die Zuordnung der Items zu den dahinterliegenden latenten Dimensionen berücksichtigt werden. In Abbildung 21 werden Items, die inhaltlich dem gleichen Teilbereich zugeordnet sind, kreisförmig dargestellt.

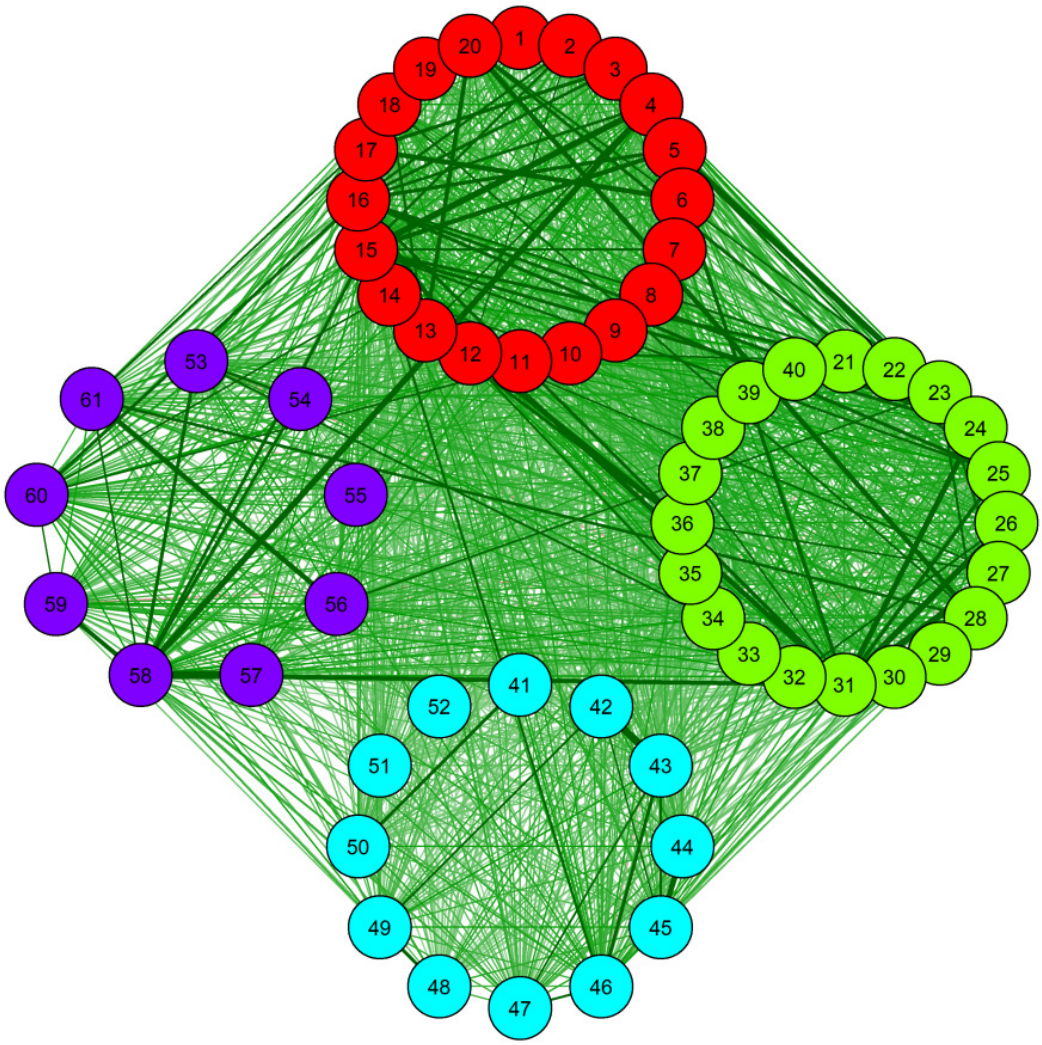


Abbildung 21 Item-Interkorrelationen als Netzwerkdarstellung mit kreisförmiger Anordnung der Items, die eine Subdimension abbilden; dicke grüne Verbindungslinien stellen hohe positive Korrelationen dar; erstellt mit dem R-Paket „Qgraph“ (Epskamp et al., 2012)

Items, die bei der Testkonstruktion dem Bereich „Service-Kultur“ zugeordnet wurden, sind in roter Farbe dargestellt, Items, die dem Konstrukt „Service-Zuverlässigkeit“ zugeordnet sind,

in grüner Farbe. Items, die dem Bereich „Umgang mit Beschwerden“ zugeordnet sind, werden in Türkis dargestellt. Die Items, die zum Bereich „Mitarbeiterqualifikation“ gehören, sind in Violett dargestellt. Diese Darstellung zeigt deutlich, dass die Items, die einem Bereich zugeordnet sind, miteinander hoch korrelieren, wobei hier auch deutlich zu erkennen ist, dass auch Items, die inhaltlich verschiedenen Bereichen zugeordnet sind, erheblich miteinander korrelieren.

Die Korrelation zwischen den Indikatoren für den Erfolg der Organisation und der Kundenzufriedenheit stellt Tabelle 4 dar. Weil die Befragung nicht von allen Versuchspersonen vollständig ausgefüllt wurde, wurde zur Bestimmung der Korrelationen ein paarweiser Fallausschluss gewählt, der dazu führt, dass die Korrelationen auf einem Stichprobenumfang zwischen N = 604 und N = 778 Personen basieren.

Tabelle 4 Korrelation der Indikatorvariablen für Kundenzufriedenheit und Erfolg der Organisation

Variable	1	2	3	4	5
1 Umsatz	-				
2 Gewinn	.61**	-			
3 Stabilität der Kundenbeziehung	.26**	.26**	-		
4 Weiterempfehlungsbereitschaft	.23**	.22**	.61**	-	
5 Wiederkaufbereitschaft	.23**	.27**	.60**	.67**	-

* p < .05, ** p < .01

Bedingt durch den hohen Stichprobenumfang unterscheiden sich alle Korrelationskoeffizienten in Tabelle 4 hoch signifikant von $r = 0$. Es zeigt sich erwartungskonform eine hohe Korrelation zwischen Umsatz und Gewinn, den Indikatoren für den Erfolg der Organisation. Auch die Indikatoren für Kundenzufriedenheit: Stabilität der Kundenbeziehung, Weiterempfehlungsbereitschaft und Wiederkaufbereitschaft korrelieren hoch miteinander. Betrachtet man die restlichen Korrelationen in Tabelle 4, so sieht man mittlere Zusammenhänge, die für einen Zusammenhang zwischen der Kundenzufriedenheit und dem Erfolg einer Organisation sprechen.

5.2 Konfirmatorische Faktorenanalyse

Welche Dimensionen bzw. latente Konstrukte sich in den empirischen Antwortdaten und deren Varianz-Kovarianz-Matrix finden lassen, wurde faktorenanalytisch untersucht. Weil es theoretische Annahmen gab, wie sich die einzelnen Items latenten Dimensionen zuordnen lassen, wurde eine konfirmatorische Faktorenanalyse durchgeführt. Es wurde zunächst ein

klassisches Modell der konfirmatorischen Faktorenanalyse berechnet (CFA), bei dem jedes Item der jeweiligen Dimension zugeordnet ist (s. Abb. 14 a). Da die vier Subfacetten gemeinsam als Indikatoren für Service-Qualität betrachtet werden, wurden die Korrelationen dieser latenten Dimensionen geschätzt. Weil Service-Qualität als ein Konstrukt zweiter Ordnung gesehen werden kann, wurde, wie in Abbildung 14 b dargestellt, zusätzlich ein Modell mit einem Faktor zweiter Ordnung geschätzt und berichtet (SOFA / Faktor zweiter Ordnung). Zum Vergleich mit diesen beiden Modellen wurde ein Modell, in dem alle Items auf einem Faktor laden, erstellt und berichtet (1 Faktor). Die Indikatoren für die Güte dieser drei Modelle stellt Tabelle 5 dar.

Tabelle 5 Modellfit Indikatoren für die Modelle der Konfirmatorischen Faktorenanalyse

Modell	$\chi^2_{(df)}$	$p_{(\chi^2)}$	χ^2/df	CFI	TLI	SRMR	RMSEA	90 % – CI
CFA	3366.56 (1704)	< .01	1.98	0.907	0.903	0.076	0.051	0.049 – 0.054
Faktor 2. Ordnung	3369.93 (1706)	< .01	1.97	0.907	0.903	0.076	0.051	0.049 – 0.054
1 Faktor	4118.85 (1710)	< .01	2.41	0.865	0.860	0.085	0.062	0.059 – 0.064

CFI = Confirmatory fit index, TLI = Tucker-Lewis index, SRMR = standardized root mean square residual, RMSEA = Root-mean-square error of approximation, 90 % CI = 90 % Konfidenzintervall für RMSEA
Modelle: „CFA" = Konfirmatorische Faktorenanalyse; „Faktor 2. Ordnung" = CFA Modell mit einem Faktor 2. Ordnung (SOFA); 1 Faktor = CFA Modell, in dem alle Items auf einem gemeinsamen Faktor laden

Der χ^2-Test fällt für alle in Tabelle 5 berichteten Modelle signifikant aus und die Nullhypothese, dass sich die empirische Varianz-Kovarianz-Matrix von der modellimplizierten Varianz-Kovarianz-Matrix nicht signifikant unterscheidet, muss zurückgewiesen werden. Wie in Kapitel 4.4 beschrieben, weist dieser Test bei großen Stichproben bereits sehr kleine und ggf. unbedeutende Abweichungen als signifikant aus. Das Verhältnis von χ^2 zu den Freiheitsgraden des CFA-Modells und des Modells mit dem Faktor zweiter Ordnung, das jeweils < 2 ausfällt, kann als Indikator für eine gute Modellpassung dieser beiden Modelle betrachtet werden. Das minimal bessere Verhältnis aus dem χ^2-Wert und den Freiheitsgraden für das Modell mit dem Faktor zweiter Ordnung hängt damit zusammen, dass dieses Modell vier Ladungskoeffizienten zwischen dem Faktor zweiter Ordnung (Service-Qualität) und den Subdimensionen schätzt. Es ist verglichen mit dem CFA-Modell, in dem sechs Interkorrelationen zwischen den vier Subdimensionen geschätzt werden, sparsamer und weist zwei Freiheitsgrade mehr auf. Das Verhältnis aus dem χ^2-Wert und den Freiheitsgraden fällt für das Modell mit nur einem Faktor mit 2.42 > 2 nicht so gut aus, was darauf hinweist, dass sich die Passung dieses Modells vergleichen mit den anderen beiden als schlechter erweist. Die weiteren Parameter zur Beurteilung der Modellgüte, deren üblichen Grenzwerte in Kapitel 4.4 dargestellt wurden, weisen auf eine

sehr gute oder gute Passung für das CFA-Modell und das Modell mit einem Faktor zweiter Ordnung hin. Auch diese Modellgüteindikatoren bestätigen, dass das Modell mit nur einem Faktor eine schlechtere Passung aufweist.

Die in Tabelle 6 berichteten Interkorrelationen der vier Facetten der Skala zur Erfassung von Service-Qualität fallen alle hoch und positiv aus. Alle geschätzten Korrelationen erweisen sich als signifikant und deuten darauf hin, dass die Annahme von Service-Qualität aus gemeinsamem Faktor, der diese Zusammenhänge erklärt, plausibel ist.

Tabelle 6 Interkorrelation der vier Subdimensionen der Skala Service-Qualität aus der Konfirmatorischen Faktorenanalyse

Dimension	1	2	3	4
1 Service-Kultur	-			
2 Service-Zuverlässigkeit	.83**	-		
3 Umgang mit Beschwerden	.76**	.66**	-	
4 Mitarbeiterqualifikation	.86**	.82**	.74**	-

* p < .05, ** p < .01

In Tabelle 11 (s. Anhang C) wird neben den Item-Namen und dem Wortlaut der Items die Zuordnung der Items zu den vier latenten Dimensionen und die standardisierten Ladungskoeffizienten dargestellt. Es zeigen sich überwiegend hohe und signifikante Ladungskoeffizienten. Lediglich Frage 40 weist einen Koeffizienten < .40 auf. Betrachtet man Abbildung 19, in der das Korrelationsmuster zwischen den Items durch räumliche Anordnung abgebildet wird, so wird an der Position des Items deutlich, dass dieses Ergebnis erwartungstreu ist. Wie in Kapitel 4.4 beschrieben, wurde bei den Items von einem ordinalen Skalenniveau ausgegangen und deshalb der WLSMV-Schätzer genutzt (Heflin et al., 2009; Li, 2016a). Die Schwellen, die zur Berechnung der polychorischen Korrelationen genutzt wurden, werden im Anhang D in Tabelle 12 dargestellt.

5.3 Test des Rahmenmodells

Die zentralen Hypothesen dieser Studie wurden als überidentifiziertes Strukturgleichungsmodell spezifiziert und geschätzt. Alle Pfadkoeffizienten des Rahmenmodells, die sich als statistisch signifikant erwiesen, werden in Abbildung 22 dargestellt.

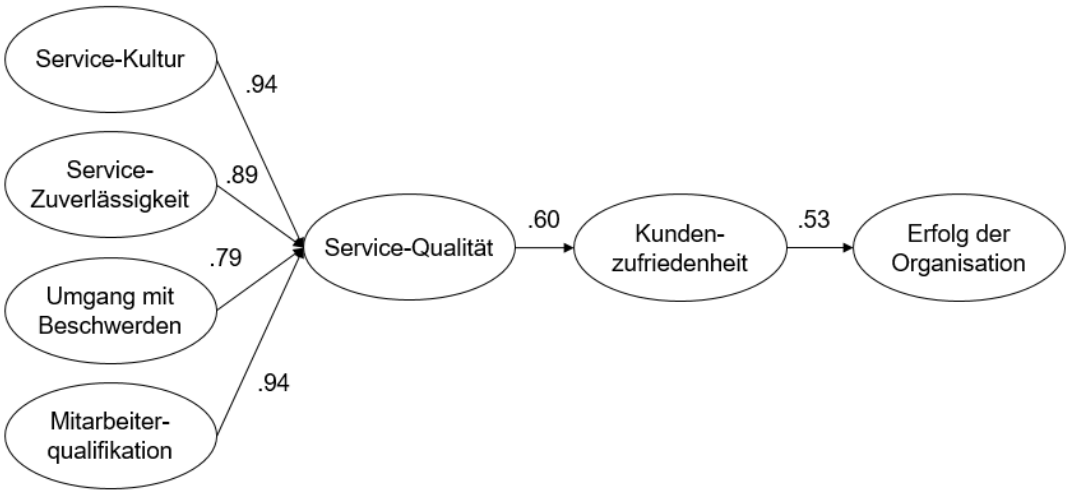


Abbildung 22 Service- Profit Chain Modell zur Vorhersage des Erfolgs einer Organisation

Zur Beurteilung der Modellgüte werden in Tabelle 7 verschiedene Fit-Indizes zusammengefasst. Die Interpretation dieser Modellgütemaße weist insgesamt auf eine gute Passung hin.

Tabelle 7 Fit Indizes für das Service-Profit Chain Modell

Model	$\chi 2_{(df)}$	$p_{(\chi 2)}$	$\chi 2/df$	CFI	TLI	SRMR	RMSEA	90 % – CI
SPC	3390.34 (2009)	< 0.010	1.69	0.980	0.979	0.077	0.047	0.044 – 0.050

CFI = Confirmatory fit index, TLI = Tucker-Lewis index, SRMR = standardized root mean square residual, RMSEA = Root-mean-square error of approximation, 90 % CI = 90 % Konfidenzintervall für RMSEA, SPC = Service-Profit Chain

Die vier Subfacetten von Service-Qualität laden alle hoch auf der latenten Variablen Service-Qualität. Dieses Ergebnis passt gut zu den Ergebnissen aus Tabelle 6, die die hohe Interkorrelationen der vier Subdimensionen aus der konfirmatorischen Faktorenanalyse darstellt. Im rechten Teil des Modells wird der Zusammenhang zwischen Service-Qualität, Kundenzufriedenheit und dem Erfolg der Organisation dargestellt. Die standardisierten Pfadkoeffizienten zeigen hier, dass der Erfolg der Organisation signifikant durch die Kundenzufriedenheit beeinflusst wird, wobei 29 % der Varianz des Erfolgs erklärt werden konnte (R^2 = .29). Zudem besteht ein substanzieller Zusammenhang zwischen Service-Qualität und Kundenzufriedenheit. Das Modell konnte die Varianz der Kundenzufriedenheit zu 36 Prozent aufklären (R^2 = .36).

6 Diskussion

Im Diskussionsteil des ersten Teils dieser Arbeit werden zunächst die wichtigsten Ergebnisse zusammengefasst. Nach der anschließenden Diskussion methodischer Einschränkungen werden Implikationen für die Praxis und Anwendungsmöglichkeiten beschrieben. Der Diskussionsteil endet mit einem Ausblick, der die Konsequenzen für die theoretische Weiterentwicklung und weitere Forschung auf diesem Gebiet aufzeigt.

6.1 Zusammenfassung der Ergebnisse

Das in Kapitel 3.4 vorgestellte Rahmenmodell dieser Studie, das die zentralen Hypothesen in einem Strukturgleichungsmodell zusammenfasst, konnte anhand der empirischen Daten bestätigt werden. Die Ergebnisse der Studie sprechen für die Konzeption von Service-Qualität als ein Konstrukt höherer Ordnung, das auf vier Subdimensionen basiert. Es wurde gezeigt, dass Service-Qualität, wie angenommen, im Zusammenhang mit Kundenzufriedenheit steht. Ebenso besteht ein Zusammenhang zwischen Kundenzufriedenheit und dem Erfolg der Organisation. Die ermittelten Pfadkoeffizienten für diese Zusammenhänge fielen hoch aus und der Anteil der erklärten Varianz in der Kundenzufriedenheit und dem Erfolg der Organisation weist darauf hin, dass das vorgeschlagene Modell einen beachtlichen Teil der Varianz dieser Variablen erklären kann. Diese Befunde passen sehr gut zu dem in Kapitel 3.3.5 beschriebenen Stand der Forschung zum Thema Service-Qualität. Das in der Literatur als Service-Profit Chain bezeichnete Modell, das davon ausgeht, dass der vielfach gefundene bivariate Zusammenhang zwischen Service-Qualität und dem Erfolg von Organisationen durch Kundenzufriedenheit vermittelt wird, konnte auch in dieser Studie empirisch bestätigt werden (Heskett et al., 1997; Hogreve et al., 2017; Hong et al., 2013). Alternative Modelle, die neben der Mediation auch einen direkten Effekt von Service-Qualität auf den Erfolg der Organisation berücksichtigten, weisen keinen besseren Modellfit auf und zeigen, dass dieser direkte Effekt klein ausfällt. Aus diesem Grund und da wissenschaftliche Erklärungsmodelle möglichst einfach und sparsam sein sollten, wird das Modell der Service-Profit Chain dem komplexeren Modell, das auch den direkten Zusammenhang von Service-Qualität und dem Erfolg der Organisation umfasst, vorgezogen (Epstein, 1984).

6.2 Methodische Einschränkungen

Es ist gelungen, eine relativ große Stichprobe zur Teilnahme an dieser Studie zu motivieren, so dass die angestrebten Strukturgleichungsmodelle berechnet werden konnten. Für die Teilnahme an dieser Studie kamen nur Personen in Frage, die einen ausreichend guten Überblick über die Service-Prozesse in ihrer Organisation haben, weshalb zur Rekrutierung auf eine Datenbank des TÜV Süd zurückgegriffen wurde. Diese Rekrutierungsstrategie könnte dazu geführt haben, dass sich die gewonnene Stichprobe in entscheidenden Merkmalen von der Population deutscher Unternehmen unterscheidet und damit nicht notwendigerweise repräsentativ für alle Unternehmen in Deutschland ist. Wie bei allen Studien, bei denen die Teilnahme freiwillig ist, kam es zu einer Selbstselektion der teilnehmenden Personen. Ungefähr 15 % der angeschriebenen Personen haben sich für die Teilnahme an der Studie entschieden. Diese Entscheidung könnte mit für diese Studie relevanten Merkmalen konfundiert gewesen sein. So wäre es zum Beispiel denkbar, dass Personen aus Organisationen, in denen das Thema Service-Qualität einen hohen Stellenwert hat, sich häufiger für die Teilnahme entschieden haben als andere. Der Aspekt der Selbstselektion und die damit möglicherweise einhergehende eingeschränkte Repräsentativität der Stichprobe sollte deshalb bei der Interpretation und vor allem bei der Generalisierung der Ergebnisse beachtet werden.

Da die Aussagekraft einer wissenschaftlichen Studie stark von der Qualität der Operationalisierung ihrer Konstrukte abhängt, sollten die in dieser Studie gewählten Operationalisierungen kritisch hinterfragt werden. Für das Konstrukt Service-Qualität existieren, wie in Kapitel 3.3.3 und 3.3.4 zusammengefasst, unterschiedliche Definitionen und Operationalisierungen. Da die bestehenden Instrumente zur Erfassung von Service-Qualität für die vorliegende Studie ungeeignet waren, wurde unter Berücksichtigung der bisherigen Arbeiten auf diesem Gebiet und gemeinsam mit einem Expertenteam ein neuer Fragebogen zur Erfassung von Service-Qualität entwickelt. Der neu entwickelte Fragebogen nutzt Items, die von verschiedenen Organisationen aus einer Vielzahl von Branchen als Selbsteinschätzung beantwortet werden können. Es ist gelungen, Service-Qualität als Konstrukt höherer Ordnung zu operationalisieren, das die vier theoretisch angenommenen Subfacetten umfasst. Sowohl die Ergebnisse der CFA als auch die Zusammenhänge mit den anderen Konstrukten weisen darauf hin, dass der gewählte Operationalisierungsansatz gelungen ist. Für die Erfassung der Kundenzufriedenheit und des Erfolgs der Organisation wurden sehr ökonomische Lösungen gewählt, die gemessen an ihrer Reliabilität als zufriedenstellend beurteilt werden können. Die Korrelationsmuster der Indikatoren dieser Konstrukte untereinander und mit den anderen Variablen dieser Studie weisen ebenfalls darauf hin, dass eine ausreichend gute Messung der Konstrukte gelungen ist.

Beurteilet man die Modellgüte des Rahmenmodells und der Modelle der CFA nach der Logik von Nullhypothesen Signifikanztests und der χ^2-Statistik, könnte man zu dem Schluss kommen, dass deren Modellgüte ungenügend ist. Eine solche Strategie wird von Gigerenzer

(2004) als gedankenlos bezeichnet, weshalb in dieser Arbeit weitere Indikatoren für die Modellgüte betrachtet wurden, die insgesamt auf eine gute Modellgüte hinweisen.

6.3 Zukünftige Forschung und theoretische Weiterentwicklung

Die bisherige Forschung zur Service-Profit Chain und auch die vorliegende Studie legen nahe, dass es sich bei den untersuchten Zusammenhängen zwischen Service-Qualität, Kundenzufrieden und dem Erfolg von Organisationen um kausale Wirkmechanismen handelt. Die eingesetzten Methoden, die die relevanten Konstrukte nur zu einem Messzeitpunkt erfassen und keine experimentelle Manipulation der Prädiktorvariablen umfassen, sind nicht in der Lage, die Frage der Kausalität ausreichend zu beantworten. Um diese Frage zu untersuchen, sollten in weiteren Studien andere Forschungsdesigns verwendet werden. Eine Möglichkeit zur Prüfung der Kausalität bestünde darin, eine experimentelle Variation der Service-Qualität umzusetzen. Diese Strategie wird für Organisationen, die sich in im Wettbewerb mit anderen Organisationen befinden, schwer umsetzbar sein. Alternativ könnte die Frage der Kausalität durch eine Zeitreihenuntersuchung, die die Erfassung der Konstrukte zu mehreren Messzeitpunkten erfordert, behandelt werden.

Für eine wiederholte Erfassung der Konstrukte wäre eine Steigerung der Ökonomie der Skala zur Erfassung von Service-Qualität wünschenswert. Dies könnte über die Entwicklung einer Kurzskala realisiert werden. Ebenso wäre es nützlich, mehrere parallele Fragebogenversionen zu entwickeln, so dass es bei den Versuchspersonen bei wiederholter Messung nicht zu Erinnerungs- oder Konsistenzeffekten kommt.

In dieser Studie wurde Service-Qualität bewusst aus der Perspektive von Mitgliedern einer Organisation betrachtet. Verzichtet wurde etwa auf die Perspektive von Kundinnen und Kunden. Um ein vollständigeres Bild zu den Annahmen dieser Arbeit zu bekommen, wäre es erforderlich, Service-Qualität und die anderen Konstrukte des Rahmenmodells auch aus anderen Perspektiven zu erfassen. Im Bereich Service-Qualität wurde mit dem GAP-Modell bereits ausführlich untersucht, wie die Perspektive der Organisation und die der Personen, die die Produkte oder Dienstleistungen der Organisation nutzen, zusammenhängen (Ali & Raza, 2017; Cronin Jr & Taylor, 1994; Parasuraman et al., 1988). Eine Erweiterung des Rahmenmodells dieser Studie mit den Perspektiven weiterer Personengruppen würde die Entwicklung einer Skala zur Erfassung von Service-Qualität aus beispielsweise der Perspektive von Kundinnen und Kunden voraussetzen. Da diese Zielgruppe interne Service-Prozesse jedoch meist nicht einschätzen kann, stellt sich hier die Aufgabe zu definieren und zu operationalisieren, welche Aspekte von Service-Qualität aus der Perspektive von Kundinnen und Kunden sinnvoll eingeschätzt werden können. Durch die Kombination der beiden Perspektiven könnten sich neue und umfassendere Erkenntnisse ergeben.

Ein solches erweitertes Forschungsdesign sollte berücksichtigen, dass, sobald mehrere Personen die Service-Qualität einer Organisation einschätzen, hierarchisch strukturierte Daten erfasst werden, die dann durch Mehrebenenmodelle ausgewertet werden können (Snijders & Roel, 2012). Dieser Ansatz könnte genutzt werden, um zu untersuchen, wie Variablen, die auf der Ebene der Organisation angesiedelt sind bzw. von Organisation zu Organisation variieren – wie z. B. deren Größe, die Branche oder das Budget, dass für Weiterbildungsangebote zum Thema Service-Qualität zur Verfügung steht –, in das Rahmenmodell integriert werden können. Ein Beispiel für ein solches Design lieferten Schuh, Egold und van Dick (2012), die unter anderem untersuchten, wie die Identifikation mit der Organisation von Führungskräften und Angestellten mit der Kundenzufriedenheit zusammenhängen.

Neben der Selbsteinschätzung der Service-Qualität durch Mitglieder einer Organisation könnte die Fremdeinschätzung durch externe Personen, die hohe Expertise im Bereich Service-Qualität besitzen, interessante Informationen liefern. Würden die relevanten Konstrukte in weiteren Studien von mehreren Personengruppen eingeschätzt, könnte der Multitrait-Multimethod-Ansatz zur Auswertung genutzt werden (Campbell & Fiske, 1959; Mahlke et al., 2016).

Um einen solchen Ansatz zu verfolgen, wäre es wünschenswert, darauf zu achten, dass die einzelnen Konstrukte ähnlich operationalisiert werden. In dieser Studie, in der der Fokus unter anderem darauf lag, ein neues Verfahren zur Erfassung von Service-Qualität zu entwickeln, wurde das Konstrukt Service-Qualität sehr breit und mit vielen Indikatoren erfasst. Die Operationalisierung von Kundenzufriedenheit und Erfolg der Organisation wurde sehr ökonomisch gestaltet. Für künftige MTMM-Studien wird empfohlen, die Konstrukte auf ähnlichem Abstraktionsniveau zu erfassen. Um noch genauer einschätzen zu können, unter welchen Bedingungen das postulierte Rahmenmodell dieser Studie gültig ist, sollten zudem zusätzliche Variablen erfasst werden. So könnte es zum Beispiel relevant sein, ob die untersuchte Organisation primär Produkte herstellt und vertreibt oder ob der Umsatz der Organisation primär durch Dienstleistungen generiert wird.

Um den Organisationen die Interpretation der Ergebnisse zur eigenen Service-Qualität zu erleichtern, wäre es hilfreich, Normwerte für verschiedene Branchen und Organisationsgrößen anzubieten. Solche Normwerte können durch die Erfassung einer großen Eichstichprobe gewonnen werden und dazu dienen, das Ergebnis einer Organisation im Vergleich mit Normwerten zu beurteilen.

6.4 Anwendungsmöglichkeiten und Ausblick

In dieser Studie ist es gelungen, eine Skala zu Erfassung von Service-Qualität zu entwickeln und einzusetzen sowie Daten zu erfassen, die zur Überprüfung der Reliabilität und Validität der Skala genutzt werden konnten. Die resultierende Skala ermöglicht die Service-Qualität einer

Organisation aus der Binnenperspektive der Organisationsmitglieder zu diagnostizieren. Bei der Entwicklung der Skala wurde darauf geachtet, dass die Items nicht spezifisch auf eine Organisation oder Branche zugeschnitten sind, weshalb ein sehr breites Anwendungsfeld möglich ist. Dies ist insbesondere für die Forschung zum Thema Service-Qualität hilfreich und ermöglicht künftige Studien, die sowohl den Vergleich verschiedener Organisationen und Branchen also auch die Untersuchung von Einflussgrößen, die auf der Ebene von Organisationen angesiedelt sind, etwa deren Kultur, vornehmen.

Die Items der Skala wurden so entwickelt, dass sie die vier theoretisch hergeleiteten Subdimensionen abdecken. Die Zuordnung der Items zu der jeweiligen Dimension konnte empirisch bestätigt werden. In der Praxis kann für einen schnellen Überblick der Gesamtwert für Service-Qualität genutzt werden. Zudem können die Werte der Subdimensionen und die Antwortverteilung zu den einzelnen Items genutzt werden, um gezielt Verbesserungspotenziale zu identifizieren.

Da die entwickelten Items alle wichtigen Bereiche des Service-Managements abdecken, können sie neben der Erfassung von Service-Qualität auch als Checkliste für Service-Qualität genutzt werden. Anhand dieser Checkliste kann geprüft werden, ob alle relevanten Elemente von Service-Qualität in einer Organisation ausreichend beachtet werden. Wenn alle Elemente der Checkliste in einer Organisation ausreichend berücksichtig werden, kann davon ausgegangen werden, dass die resultierende Service-Qualität hoch ist.

Soll die entwickelte Skala zur Erfassung von Service-Qualität in wissenschaftlichen Studien in Verbindung mit anderen Befragungsinstrumenten eingesetzt werden, könnte argumentiert werden, dass die Ökonomie des Verfahrens aufgrund der hohen Anzahl an Items und der damit einhergehenden langen Bearbeitungszeit relativ gering ist. Auch für eine regelmäßige Einschätzung der Service-Qualität im Rahmen des Service-Managements in einer Organisation wäre eine kürzere Bearbeitungsdauer vorteilhaft. Die Entwicklung einer Kurzskala wäre eine Möglichkeit, um die Bearbeitungsdauer zu reduzieren. Eine geringere Anzahl von Fragebogen-Items einer solchen Kurzskala wäre mit einer geringeren Reliabilität und gegebenenfalls einer Einschränkung der Konstruktvalidität verknüpft. Eine weitere Option, die Bearbeitungsdauer der Skala zu minimieren, ohne dabei eine geringere Reliabilität der Messung in Kauf zu nehmen, besteht in der Entwicklung eines adaptiven Tests zur Erfassung von Service-Qualität. Wie ein adaptiver Test zur Erfassung von Service-Qualität entwickelt werden kann und welche Entscheidungen dabei berücksichtigt werden sollten, wird im zweiten Teil dieser Arbeit thematisiert.

7 Entwicklung eines adaptiven Tests zur Erfassung von Service-Qualität

Im zweiten Teil dieser Arbeit wird einleitend die Grundidee des adaptiven Testens dargestellt. Nach einer systematischen Einordnung verschiedener Varianten adaptiver Tests wird zusammengefasst, welche Argumente für und welche gegen die Nutzung adaptiver Tests in der diagnostischen Praxis sprechen. Anschließend wird der prototypische Ablauf adaptiver Tests dargestellt und beschrieben, welche Optionen bei der Entwicklung der einzelnen Teilschritte gegeben sind. Um herauszufinden, welche Konfiguration für einen adaptiven Test zur Erfassung von Service-Qualität zu den besten Ergebnissen führt, werden anschließend in einer Simulationsstudie verschiedene Konfigurationsoptionen verglichen. Die Ergebnisse der Simulationsstudie werden diskutiert und für die praktische Entwicklung des adaptiven Tests genutzt. Abschließend wird beschrieben, welche technischen Grundlagen entwickelt wurden und wie die entwickelten Elemente in einen webbasierten adaptiven Test zur Erfassung von Service-Qualität integriert wurden.

7.1 Einleitung: Grundidee des adaptiven Testens

Die meisten psychodiagnostischen Verfahren bestehen aus einer festen Anzahl von Testaufgaben, die von allen Teilnehmenden in fester Reihenfolge bearbeitet werden sollen. Adaptive Tests passen sich, wie der Name bereits vermuten lässt, während der Testung an das Antwortverhalten und die daraus geschätzte Merkmalsausprägung der Testperson an. Führt man sich diese Besonderheit vor Augen, könnte man vermuten, dass es sich bei adaptiven Tests um eine revolutionäre neue Idee handelt. Mit etwas Abstand betrachtet, stellt man fest, dass eine schrittweise Annäherung und das immer genauere Einkreisen eines Ergebnisses eine Strategie darstellen, die Menschen im Alltag häufig einsetzen. Möchte man zum Beispiel erfahren, wie zufrieden jemand mit seiner zurückliegenden Urlaubsreise ist, so würde man zunächst mit einer allgemeinen Frage beginnen. Würde man bereits auf diese erste Frage erfahren, dass die befragte Person sehr unzufrieden mit der Reise war, so würde man im Anschluss Fragen stellen, die dazu dienen herauszufinden, welche Faktoren zu dieser Unzufriedenheit geführt haben. Die Frage danach, ob man diese Urlaubsreise weiterempfehlen oder wiederholen würde, erübrigt sich in diesem Fall, denn die erwartbare Antwort – ein klares Nein – kann bereits aus der Antwort auf die erste Frage abgeleitet werden.

Auch eine mündliche Prüfung stellt eine Situation dar, in der die Grundidee des adaptiven Testens gut beobachtet werden kann. Beantwortet ein Prüfling eine Reihe leichter Prüfungs-

fragen richtig, wird davon ausgegangen, dass die Person ein gewisses Maß an Kompetenz aufweist, und es werden in der Folge schwierigere Prüfungsfragen gestellt. Scheitert ein Prüfling wiederholt an sehr schweren Fragen, so werden im weiteren Verlauf leichtere Prüfungsfragen genutzt. Dem Prüfenden ist in einem solchen Szenario klar, dass es keinen weiteren Informationswert hat, weiterhin schwere Fragen zu stellen, denn es kann davon ausgegangen werden, dass diese Fragen ebenfalls nicht zufriedenstellend beantwortet werden können.

Versucht man, den Prozess, der unter anderem einer mündlichen Prüfung zugrunde liegt, genauer nachzuvollziehen, wird deutlich, dass zunächst Wissen über mögliche Prüfungsfragen benötigt wird. Zusätzlich zum Wortlaut dieser Prüfungsfragen wird eine Annahme darüber benötigt, wie schwer die potenziellen Prüfungsfragen zu beantworten sind. Darüber hinaus muss aufgrund der Antworten abgeschätzt werden können, wie gut die Leistung eines Prüflings in etwa ist. Bei der Suche nach einer geeigneten nächsten Prüfungsfrage sollte – ausgehend von der Annahme, wie gut der Prüfling ist, und dem Wissen über die Schwierigkeit aller potenziellen weiteren Fragen – eingeschätzt werden, welche nächste Frage nützlich ist, um zu einer noch präziseren Einschätzung der Prüfungsleistung zu gelangen.

7.2 Varianten adaptiver Tests

Adaptive Tests lassen sich anhand verschiedener Merkmale unterscheiden. Es gibt solche, die insofern adaptiv sind, als sie anhand von im Handbuch festgelegten Verzweigungsregeln definieren, welche Aufgabe eine Person nach dem erfolgreichen oder nicht erfolgreichen Lösen einer aktuellen Aufgabe als nächstes bearbeiten soll. Wegen ihres verzweigten Aufbaus werden diese Tests in der englischsprachigen Fachliteratur „branched tests“ genannt. Sie haben den Vorteil, dass bei der Durchführung kein Computer benötigt wird. Gleichzeitig setzen sie voraus, dass die Person, die den Test durchführt, mit den Verzweigungsregeln gut vertraut ist. Ein Beispiel für ein solches Testverfahren ist z. B. das Adaptive Intelligenz Diagnostikum (AID) von Kubinger und Wurst (1985) in seiner ersten Auflage. In der überarbeiteten zweiten Version (AID 2) wurde für die meisten Testteile an der Idee des verzweigten Testens festgehalten. Die Schwierigkeit der Aufgaben wird von der Testleitung an das Leistungsniveau der Testperson in den vorangegangen Aufgabengruppen angepasst (Jacobs, Heubrock & Petermann, 2003). Auch andere Testverfahren aus dem Bereich der Diagnostik kognitiver Leistungsfähigkeit, wie der WAIS VI, nutzen solche Verzweigungsregeln in Verbindung mit Abbruchregeln, um mit möglichst wenig Testaufgaben eine möglichst präzise Diagnose zu ermöglichen (Benson, Hulac & Kranzler, 2010; Canivez, 2013).

Zu den in der englischen Fachliteratur sogenannten „tailored tests“ gehören Tests, bei denen das Fähigkeits- oder Leistungsniveau, das in der Sprache der Testtheorie auch als Personenparameter bezeichnet wird, nach jeder Aufgabe neu approximiert wird. Anhand dieser Schätzung

wird mithilfe eines Algorithmus bestimmt, welche Testaufgabe als nächstes vorgegeben werden soll. Die Bestimmung des Personenparameters und die Algorithmen zur Selektion des nächsten Items umfassen aufwendige Rechenoperationen. Diese können von der Versuchsleitung in der Regel nicht ohne Unterstützung von Computerprogrammen durchgeführt werden. Kubinger und Spohn (2017) bieten für den AID in seiner dritten Auflage ein Computerprogramm an, das die Testleitung bei genau dieser Aufgabe unterstützt. Durch diese Art der computerunterstützten Selektion des nächsten Items wird die Testdauer im Vergleich zu der bisherigen verzweigten Darbietung ungefähr halbiert. Die Ökonomie des Verfahrens steigt deutlich und der energetisch-emotionale Aufwand für die Versuchsperson ist geringer (Kubinger, 2017).

Bei den bislang vorgestellten Verfahren werden die Testmaterialien primär papierbasiert dargeboten und bearbeitet. Bei Verfahren, die als Computerbasierte Adaptive Tests (CAT) bezeichnet werden, erfolgt die gesamte Erhebung am Computer. Aus dem vorliegenden Antwortvektor wird ein vorläufiger Personenparameter geschätzt (Kapitel 7.4.3), der als Grundlage zur Bestimmung des nächsten Items durch einen Selektionsalgorithmus (Kapitel 7.4.5) genutzt wird (Choi, Reise, Pilkonis, Hays & Cella, 2010; Han, 2018; van der Linden, Wim J., 2008). Sowohl die Instruktionen als auch die Testaufgaben und die Antworten oder Lösungen werden von der Versuchsperson direkt am Computer bearbeitet. Zwei Metaanalysen untersuchten, ob die computerbasierte Darbietung von Testverfahren, verglichen mit der Darbietung auf Papier, negative Auswirkungen auf die psychometrische Qualität der Messung haben (Finger & Ones, 1999; Mead & Drasgow, 1993). Die Ergebnisse beider Metaanalysen weisen darauf hin, dass die Darbietung und Bearbeitung der Testverfahren am Computer keinen wesentlichen Einfluss auf die psychometrische Messqualität haben. Da bei computerbasierten adaptiven Tests in der Regel keine Testleitung mehr benötigt wird, kann der Aufwand für die Testung noch weiter reduziert werden, ohne nennenswerte Einbußen bei der Qualität der Messung befürchten zu müssen. Fehler, die Diagnostikerinnen und Diagnostiker bei der Durchführung und Auswertung von adaptiven Tests machen können, werden bei einer computerbasierten Diagnostik ausgeschlossen. Die Diagnostik mit Hilfe von Computersystemen erfolgt hochgradig standardisiert, so dass die Durchführungs- und Auswertungsobjektivität dieser Verfahren hoch ausfällt.

Ein weiteres Unterscheidungsmerkmal von adaptiven Tests hängt mit den eingesetzten Testaufgaben zusammen. Entweder stammen diese aus einer sogenannten Item-Bank bzw. einem Item-Pool oder sie werden während der Testung in Echtzeit generiert. Voraussetzung für die Generierung von Items mit einer definierten Item-Schwierigkeit ist, dass die kognitiven Prozesse, die zur korrekten Lösung einer Testaufgabe benötigt werden, bekannt sind (Irvine & Kyllonen, 2002). Ist dies der Fall, können die Elemente, die zur Item-Schwierigkeit beitragen, die sogenannten radicals, beim Generieren der Items so variiert werden, dass ein Item mit einer intendierten Schwierigkeit erstellt werden kann (Irvine, 2002). Beim Frankfurter Adaptiven Konzentrationsleistungs-Test (FAKT-II) ist die Darbietungsdauer der Items das schwierigkeits-

bestimmende Aufgabenmerkmal (Weis & Nuerk, 2011). In diesem Verfahren werden Aufgaben mit einer bestimmten Item-Schwierigkeit generiert, indem die Darbietungsdauer variiert wird. In einigen Forschungsgebieten sind die psychometrischen Eigenschaften, trotz regelbasierter Item-Generierung, nicht unbedingt a priori bekannt. In diesem Fall können entlang der angenommen Regeln der Generierung Item-Pools erstellt werden, deren psychometrische Messeigenschaften im Rahmen der Item-Kalibrierung empirisch untersucht werden (Geerlings, van der Linden, Wim J. & Glas, 2013). Sind die kognitiven Prozesse und Aufgabenmerkmale, die die psychometrischen Messeigenschaften ausmachen, noch weitgehend unerforscht und fehlen die Regeln, nach denen Test-Items in Echtzeit generiert werden können, werden, wie in Kapitel 7.4.1 beschrieben, die psychometrischen Messeigenschaften mit Hilfe von Modellen der Item-Response-Theorie (IRT) untersucht. Geeignete Items mit bekannten Item-Schwierigkeiten werden zu einem Item-Pool zusammengestellt. Während eines adaptiven Tests wird dann, wie in Kapitel 7.4.5 beschrieben, mittels eines Algorithmus anhand der bekannten Item-Schwierigkeit das nächste Item ausgewählt.

Adaptive Tests lassen sich auch darin unterscheiden, welche Inhalte sie erfassen. Ihren Ursprung haben adaptive Tests in der Leistungsdiagnostik. Frühe adaptive Tests wurden für Konstrukte wie Intelligenz, Konzentrationsleistung und schulische Leistung entwickelt. Grundsätzlich können auch psychologische Konstrukte, bei denen es nicht um maximales, sondern eher typisches Verhalten geht, mit adaptiven Tests erfasst werden. Zahlreiche Studien haben gezeigt, dass adaptive Tests bei der Messung psychologischer Konstrukte, wie Persönlichkeit, Angst, Depression, Wut und selbstmörderisches Verhalten, Sucht und emotionale Belastung, zu ausreichend genauen Testergebnissen führen (Carroll, 2013; De Beurs, de Vries, de Groot, de Keijser & Kerkhof, 2014; Fliege et al., 2005; Gibbons et al., 2012; Kirisci et al., 2012; Ortner, 2008; Pilkonis et al., 2011; Walter et al., 2007).

Der in dieser Arbeit entwickelte adaptive Test zur Erfassung von Service-Qualität wurde wegen der höheren Ökonomie als „tailored test“ konzipiert. Er ist ein computerbasierter adaptiver Test (CAT), der webbasiert dargeboten wird. Die Erfassung kann ohne eine Testleitung an jedem internetfähigen Gerät mit einem aktuellen Browser erfolgen. Da die kognitiven Prozesse, die den Urteilen über Service-Qualität zugrunde liegen, noch nicht ausreichend erforscht sind, können die Items nicht während der Testung generiert werden; es wird auf einen Pool von vorab entwickelten und kalibrierten Items zurückgegriffen.

7.3 Chancen und Risiken adaptiver Tests

In diesem Abschnitt werden die Chancen beschrieben, die die Entwicklung eines adaptiven Tests mit sich bringt. Dabei wird auf Aspekte wie Ökonomie, Akzeptanz, Motivation und Testangst eingegangen. Die zentralen wissenschaftlichen Studien zu diesen Aspekten werden

zusammengefasst und eingeordnet. Zudem werden Risiken und Herausforderungen, insbesondere aus psychometrischer Perspektive, dargestellt, die im Zusammenhang mit der Entwicklung adaptiver Tests berücksichtigt werden sollten.

7.3.1 Kosten und Ökonomie

Adaptive Tests können im Vergleich zu traditionellen linearen, nicht verzweigten Testverfahren mit 50 % der Items eine Messung liefern, deren Messgenauigkeit gleich oder besser ist (Embretson & Reise, 2000; Weiss, 1982, 2011). Diese Reduktion der Anzahl an genutzten Items geht mit einer kürzeren Bearbeitungsdauer einher, weshalb adaptive Tests verglichen mit klassischen Testverfahren im Hinblick auf das Nebengütekriterium Ökonomie als deutlich überlegen anzusehen sind. Neben der verkürzten Bearbeitungszeit zeichnen sich computerbasierte adaptive Tests dadurch aus, dass die Erfassung, Auswertung und Interpretation der Daten weitgehend automatisiert ist, so dass hohe Durchführungs-, Auswertungs- und Interpretationsobjektivität gegeben ist und der Einsatz solcher Verfahren mit geringen laufenden Kosten verbunden ist.

Im Kontext von Organisationen, deren Ziel die Gewinnerwirtschaftung ist, lohnt es sich deshalb genau zu prüfen, ob konventionelle Befragungsinstrumente, die zum Beispiel zur Mitarbeiter- oder Kundenbefragung eingesetzt werden, auch als adaptive Tests umgesetzt werden können. Gelingt es eine Mitarbeiterbefragung, die bislang 20 Minuten der Arbeitszeit in Anspruch genommen hat, mittels adaptiver Testung auf die Hälfte der Bearbeitungsdauer zu verkürzen, so sollte sich relativ leicht berechnen lassen, welche finanziellen Einsparungen dadurch für die Organisation als Ganzer möglich sind.

7.3.2 Akzeptanz, Motivation und Flow

Häufig wird bei Befragungen eine geringe Beteiligung beklagt. Die Gründe für diese Nichtbeteiligung sind vielfältig. Eine Begründung ist die hohe Bearbeitungsdauer. Hinweise drauf, dass zeitintensive Befragungen auf geringe Akzeptanz stoßen, findet sich in unvollständigen Datensätzen, die erkennen lassen, dass die Teilnehmenden die Bearbeitung vorzeitig abgebrochen haben.

Welchen Einfluss adaptives Testen auf die Motivation zur Testbearbeitung hat, wurde in vielen Studien untersucht. Zunächst wurde davon ausgegangen, dass adaptives Testen im Vergleich zu nicht-adaptiven Testen eine motivationssteigernde Wirkung hat (Betz & Weiss, 1976; Frey, Hartig & Moosbrugger, 2009). Insbesondere bei leistungsschwächeren Personen wurde eine Motivationssteigerung erwartet. Frey et al. (2009) untersuchten am Beispiel des Frankfurter Adaptiven Konzentrationsleistungs-Tests, wie sich die Motivation zur Testbearbeitung zwischen der nicht-adaptiven und der adaptiven Variante des Tests unterschieden. Ihre Ergebnisse zeigen

einen gegenteiligen Effekt. Diesen motivationsmindernden Effekt der adaptiven Testform erklären die Autoren mit der subjektiven Einschätzung der Erfolgswahrscheinlichkeit. Eine Studie von Ortner, Weißkopf und Koch (2014) stellt ebenfalls fest, dass die Motivation während der Bearbeitung einer adaptiven Version eines Matrizentests geringer war als bei der nicht adaptiven Version des Verfahrens. Ling, Attali, Finn und Stone (2017) untersuchten mit einem Verfahren zur Beurteilung von mathematischen Fähigkeiten motivationale Unterschiede zwischen der klassischen Darbietung aller Aufgaben, einer adaptiven Variante und einer adaptiven Variante, die gezielt leichtere Aufgaben darbot. Bezogen auf die Motivation fanden diese Autoren keine Unterschiede zwischen den drei Erfassungsvarianten. Die Ergebnisse zum Einfluss adaptiver Tests auf die Motivation der Versuchspersonen können insgesamt als uneindeutig interpretiert werden. Im Kontext von Leistungsbeurteilungen deutet sich an, dass adaptive Testverfahren möglicherweise motivationsmindernde Effekte haben. Eine mögliche Erklärung dafür ist, dass die Anpassung der Aufgabenschwierigkeit an die Fähigkeit der Testperson dazu führt, dass die Testperson nur ungefähr die Hälfte der gestellten Aufgaben richtig lösen kann. Welche motivationalen Effekte von adaptiven Tests bei der Erfassung von nicht leistungsbezogenen Konstrukten ausgehen, wurde bislang nicht untersucht.

Das Konstrukt Flow passt theoretisch gut zur Idee des adaptiven Testens und wurde deshalb in diesem Zusammenhang untersucht. Unter Flow versteht man das (selbst-)reflexionsfreie Aufgehen in einer glatt laufenden Tätigkeit, die man trotz hoher Beanspruchung noch unter Kontrolle hat (Csikszentmihalyi, 2014). Theoretisch sollte Flow vor allem bei einer guten Übereinstimmung zwischen der Fähigkeit der Person und der Schwierigkeit der Aufgabe zustande kommen. Rheinberg und Vollmeyer (2003) variierten die Schwierigkeitsstufen eines Computerspiels und konnten zeigen, dass mittelschwere Einstellungen verglichen mit zu leichten und zu schweren Spieleinstellungen mit mehr Flow-Erleben einhergehen. Deville (1993) weist darauf hin, dass adaptive Tests bzw. tailored testing (Begriffsklärung in Kapitel 7.2) das Phänomen Flow ausnutzen können. Die Testsituation soll dadurch angenehmer empfunden werden und die individuelle Testleistung könne verbessert werden. Durch den Vergleich einer Vorgabe aller Test-Items und einer adaptiven Testung konnten Ortner et al. (2014) keinen signifikanten Unterschied im berichteten Flow-Erleben feststellen.

7.3.3 Angst und Prüfungsangst

Ob und wie adaptive Testverfahren im Zusammenhang mit dem Erleben von Angst bzw. Prüfungsangst stehen, wurde als Forschungsfrage von verschiedenen Studien untersucht. Powers (2001) verglich die papierbasierte Version des Graduate Record Examinations Test mit einer computer-basierten adaptiven Version und konnte keine Unterschiede in der Testangst feststellen. Tonidandel, Quiñones und Adams (2002) untersuchten, wie die Schwierigkeit eines

adaptiven Tests mit Testangst zusammenhängt. Sie fanden heraus, dass dieser Zusammenhang gemäß ihrer Hypothese von der wahrgenommenen Testleistung mediiert wird. Ihre Befunde deuten darauf hin, dass höhere Testangst nur dann zu beobachten ist, wenn die Bearbeitung von schwierigen Aufgaben mit einer geringen wahrgenommenen Testleistung einhergeht. Die Ergebnisse von Ortner et al. (2014) weisen ebenfalls darauf hin, dass beim adaptiven Testen mehr situative Angst berichtet wird, wenn die subjektive Erfolgswahrscheinlichkeit gering eingeschätzt wird. Welchen Einfluss adaptive Tests auf Personen mit hoher Prüfungsangst haben, wurde von Ortner und Caspers (2011) untersucht. Sie konnten zeigen, dass Personen mit hoher Prüfungsangst bei adaptiver Testung geringere Testleistungen erbringen. Die Aufklärung der Versuchspersonen darüber, dass bei der adaptiven Testung die Schwierigkeit der nächsten Aufgabe davon abhängig ist, ob die letzte Aufgabe richtig oder falsch gelöst wurde, führte zu besseren Testergebnissen.

Insgesamt deuten die vorliegenden Studien darauf hin, dass adaptive Tests verglichen mit nicht-adaptiven Tests das Potenzial haben, mehr Angst auszulösen. Studien, die diese Zusammenhänge genauer betrachteten, zeigen, dass die Wahl verschiedener Gestaltungsmerkmale das Angsterleben beeinflussen (Ling et al., 2017; Ortner & Caspers, 2011; Tonidandel et al., 2002). Werden zum Beispiel aufklärende Instruktionen gegeben oder gezielt einfachere Aufgaben präsentiert, kann die Angst reduziert werden (Ling et al., 2017; Ortner & Caspers, 2011).

7.3.4 Item-Abfolge und Reihenfolgeeffekte

Aus der Forschung zu Sequenz-, Übertragungs- und Kontexteffekten ist bekannt, dass die Abfolge, in der die Items eines diagnostischen Instruments dargeboten werden, Einfluss auf das Antwortverhalten der Versuchspersonen haben kann (Franke, 1997; Ortner, 2008; Rost & Hoberg, 1997). Bei nicht adaptiven diagnostischen Verfahren ist die Abfolge der Items normalerweise für alle Teilnehmenden gleich, wodurch die beschriebenen Sequenzeffekte bei jeder Person und Messung gleichermaßen auftreten. Die damit einhergehende Verzerrung der Messung der wahren Merkmalsausprägung ist systematischen und für alle Testungen konstant. Bei adaptiven Tests ist die Auswahl und Abfolge der Items vom Antwortverhalten der Testperson abhängig. Dies kann zu unbekannten und unbeabsichtigten Reihenfolgeeffekten führen, die für jede Person und Testung unterschiedlich ausfallen können. Dieser nicht standardisierte Messverlauf kann dazu führen, dass unsystematische Reihenfolgeeffekte die Reliabilität und Validität der Messung mindern.

Ortner (2004) untersuchte mittels Rasch-Modellen, ob sich die Item-Schwierigkeit verändert, wenn die Items in umgekehrter Abfolge dargeboten werden. Sie stellt fest, dass dies für einige Items der Fall war und weist darauf hin, dass dieses Ergebnis auf eine ernsthafte Herausforderung hinweist, die beim Entwickeln von adaptiven Tests berücksichtigt werden sollte.

Beim Vergleich der Darbietung von Items in einer festen mit einer zufälligen Abfolge wurden von Walter und Rose (2013) die im Anschluss bestimmten Item-Parameter und die Personenparameter gegenübergestellt. Sie stellen einen geringen Unterschied der beiden Darstellungsmodi fest, lediglich für die ersten beiden Items der linearen Darbietung zeigen sich erkennbare Unterschiede in den betrachteten Parametern, was sie mit einem „warming-up effect" erklären (Walter & Rose, 2013, S. 89).

In einer Studie von Ortner (2008) wurden drei adaptive Varianten der deutschen Version des Eysenck Personality Profiler Fragebogens erstellt, die sich darin unterscheiden, ob das erste Test-Item einer mittelmäßigen oder extremen Persönlichkeitsausprägung entspricht. Bei drei der sieben Subskalen des Fragebogens wurden durch die unterschiedlichen Versionen signifikant unterschiedliche Personenparameter geschätzt. Dieses Ergebnis unterstreicht die Relevanz des ersten Items, das bei einer adaptiven Testung dargeboten wird, und belegt, dass Kontexteffekte bei computerbasierten adaptiven Tests, die zur Messung von Persönlichkeitsmerkmalen eingesetzt werden, Einfluss auf das Testergebnis haben können.

7.3.5 Gesamtübersicht und Korrektur von Antworten

Wird ein diagnostisches Verfahren papierbasiert dargeboten, kann die Versuchsperson sich meistens einen Überblick über alle Aufgaben des Verfahrens verschaffen. Es besteht häufig die Möglichkeit, die Testaufgaben in einer selbst gewählten Reihenfolge zu bearbeiten, Aufgaben zu überspringen und gegebene Antworten zu einem späteren Zeitpunkt zu korrigieren. Insbesondere bei papierbasierten Leistungstests im schulischen Kontext werden Schülerinnen und Schüler dazu angehalten, sich zunächst einen Überblick über alle Aufgaben zu verschaffen und bei der Bearbeitung gezielt Strategien einzusetzen, mit deren Hilfe sie ihr Ergebnis verbessern können.

Bei einer adaptiven Testung werden die Items in der Regel einzeln und sequenziell präsentiert. Dadurch einfällt die Möglichkeit, alle Testaufgaben vorab zu überblicken. Bei den meisten adaptiven Verfahren ist es nicht möglich, Items zu überspringen und bereits bearbeitete Aufgaben erneut aufzurufen, um die Antworten zu revidieren. Dieser Erfassungsmodus kann für Personen mit Prüfungsangst, wie in Kapitel 7.3.3 beschrieben, nachteilig sein. Verschiedene empirische Studien deuten darauf hin, dass die Möglichkeit zur Überarbeitung der gegebenen Antworten von den Versuchspersonen geschätzt wird und Angst reduzieren kann (Kruger, Wirtz & Miller, 2005; Liu, Bridgeman, Gu, Xu & Kong, 2015). Die Korrektur von bereits gegebenen Antworten durch die Versuchsperson ist bei adaptiven Verfahren in der Regel nicht möglich. Dies hängt damit zusammen, dass anhand bestimmter Algorithmen, die in Kapitel 7.4.3 näher beschrieben werden, die Wahl des nächsten Test-Items von den bereits gegebenen Antworten abhängig ist. Eine Veränderung zurückliegender Antworten hätte einen anderen Testverlauf zur Folge (Papanastasiou & Reckase, 2007). Die Option, bereits gegebene Antworten zu kor-

rigieren, ermöglicht darüber hinaus, unerwünschte Strategien anzuwenden, mit denen bessere Testergebnisse erzielen werden können. Bei der Wainer-Strategie werden die Testaufgaben bei der ersten Bearbeitung absichtlich falsch beantworten, so dass der adaptive Test anschließend die einfachsten Aufgaben darbietet (Wainer, 1993). Beim Bearbeiten dieser einfachen Aufgaben versucht die Versuchsperson dazuzulernen. Im zweiten Schritt wird zu den schwierigeren Testaufgaben zurückgegangen, um das neue Wissen einzusetzen und möglichst alle Aufgaben des Tests korrekt zu beantworten. Die Kingsbury-Strategie basiert darauf abzuschätzen, ob das aktuell dargebotene Item schwieriger oder leichter ist als das zurückliegende. Hat die Versuchsperson den Eindruck, dass das aktuelle Item leichter als das vorherige ist, kann sie daraus schließen, dass das zurückliegende Item falsch gelöst wurde, und diese Antwort korrigieren (Kingsbury, 1996; Wise, Finney, Enders, Freeman & Severance, 1999). Einige Autorinnen und Autoren haben Lösungen entwickelt, um es trotz der beschriebenen Strategien und Schwierigkeiten zu ermöglichen, Antworten auf zurückliegende Testaufgaben zu überarbeiten (Han, 2013; Papanastasiou & Reckase, 2007; Sari & Raborn, 2018; Wang, Fellouris & Chang, 2019; Zhongmin, Chunyan, Yong & Hanwei, 2018). Für den in dieser Arbeit entwickelten adaptiven Test wurde entschieden, das Navigieren durch die Fragebogen-Items und die nachträgliche Korrektur bereits gegebener Antworten zu unterbinden.

7.4 Ablauf adaptiver Tests

Um den Ablauf eines adaptiven Tests systematisch und strukturiert darzustellen, wird er in Abbildung 23 vereinfacht mit Hilfe der vereinheitlichten Modellierungssprache (UML) abgebildet.

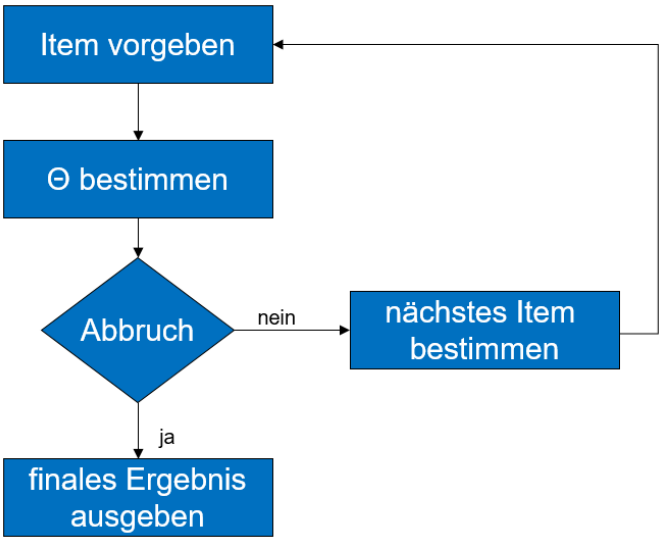


Abbildung 23 Schematische Darstellung des Ablaufs adaptiver Tests

Wie in Abbildung 23 dargestellt, besteht ein adaptiver Test aus einem Kreislauf, in dem ein Item vorgegeben wird, aus dessen Beantwortung der Personenparameter Θ bestimmt wird. Basierend auf dieser Schätzung und den Item-Parametern wird anschließend das nächste darzubietende Item bestimmt und der Versuchsperson vorgegeben. Welche Algorithmen zur Auswahl des nächsten Items eingesetzt werden können, wird in Kapitel 7.4.5 dargestellt. Dieser Kreislauf wird so lange durchlaufen, bis ein vorab definiertes Abbruchkriterium erfüllt ist. Aus der Perspektive einer Versuchsperson beginnt dieser Prozess mit der Darbietung eines ersten Items. In Kapitel 7.4.2 wird genauer darauf eingegangen, welche Überlegungen bei der Wahl des ersten Test-Items berücksichtigt werden sollten. Sobald die Antwort auf das erste Item vorliegt, wird aus dem erfassten Antwortverhalten ein vorläufiger Θ-Wert geschätzt. Bei der Schätzung dieses Wertes können grundsätzlich verschiedene Verfahren eingesetzt werden, die in Kapitel 7.4.3 genauer beschrieben werden. Neben dem Θ-Wert wird auch ein Standardfehler bestimmt, der ein Maß für die Genauigkeit der Θ-Schätzung darstellt. Da die Bestimmung des Standardfehlers an das gewählte Verfahren der Θ-Schätzung gekoppelt ist, wird in Kapitel 7.4.3 auch auf die Möglichkeiten zur Schätzung des Standardfehlers eingegangen. Nach dieser Schätzung wird entschieden, ob der Test beendet oder weitere Items vorgegeben werden sollen. Diese Entscheidung kann von verschiedenen Kriterien abhängig gemacht werden. Kapitel 7.4.4 beschreibt genauer, welche Kriterien in welcher Weise zu einem binären Abbruchkriterium zusammengefasst werden können. Wird das Abbruchkriterium erfüllt, wird die Testung beendet und die letzte Schätzung des Θ-Werts und deren Standardfehler als finales Ergebnis betrachtet. Abschließend kann das Ergebnis als Punktschätzer oder als Intervall kommuniziert werden. Neben dem reinen Ergebnis kann der Verlauf der adaptiven Testung gespeichert und gegebenenfalls der Versuchsperson zurückgemeldet werden.

7.4.1 Bestimmung der Item-Parameter

Wie in Kapitel 7.1 beschrieben wurde, ist eine Voraussetzung, um einen adaptiven Test erstellen und durchführen zu können, die Kenntnis von Aufgaben bzw. Items, aus deren Bearbeitung auf die Fähigkeit einer Person geschlossen werden kann. Die Fähigkeit einer Person wird auch als Personenparameter bezeichnet und mit dem Formelzeichen Θ_i ausgedrückt. Um im Testverlauf die Aufgaben vorzugeben, die weder zu leicht noch zu schwer sind, muss darüber hinaus die Item- bzw. Aufgabenschwierigkeit (β_j) bestimmt werden. Im Rahmen der probabilistischen Testtheorie kann die Aufgabenschwierigkeit als ein Item-Parameter anhand empirischer Daten geschätzt werden. Der klassische Ansatz der Item-Response Theorie (IRT), der seine Wurzeln in der Bildungsforschung hat, berücksichtigt zunächst nur binäres Antwortverhalten (Aufgabe gelöst, Aufgabe nicht gelöst). Die Modellgleichung des prominentesten Modells dieser Familie, des dichotomen Ein-Parameter-Rasch-Modells (1PL-Modell), wird in Formel 1 dargestellt.

$$P(U_{ij}) = 1 \mid \theta_i, \beta_j) = \frac{e^{\theta_i - \beta_j}}{1 + e^{\theta_i - \beta_i}} \quad (1)$$

Das Rasch Modell beschreibt den Zusammenhang zwischen der Fähigkeit Θ einer Person i, die auch als Personenparameter (Θ_i) bezeichnet wird, und der Wahrscheinlichkeit, eine Aufgabe richtig zu lösen ($P(U_{ij}=1)$). Da sich die Lösungswahrscheinlichkeit einer Aufgabe im Wertebereich zwischen 0 (nicht oder falsch gelöst) und 1 (richtig gelöst) bewegt, wird in Formel 1 der Zusammenhang zwischen Personenparameter und Lösungswahrscheinlichkeit als logistische Funktion spezifiziert. Abbildung 24 zeigt den Zusammenhang zwischen Personenparameter und Lösungswahrscheinlichkeit für drei Aufgaben mit unterschiedlicher Aufgabenschwierigkeit anhand von drei Funktionsgraphen.

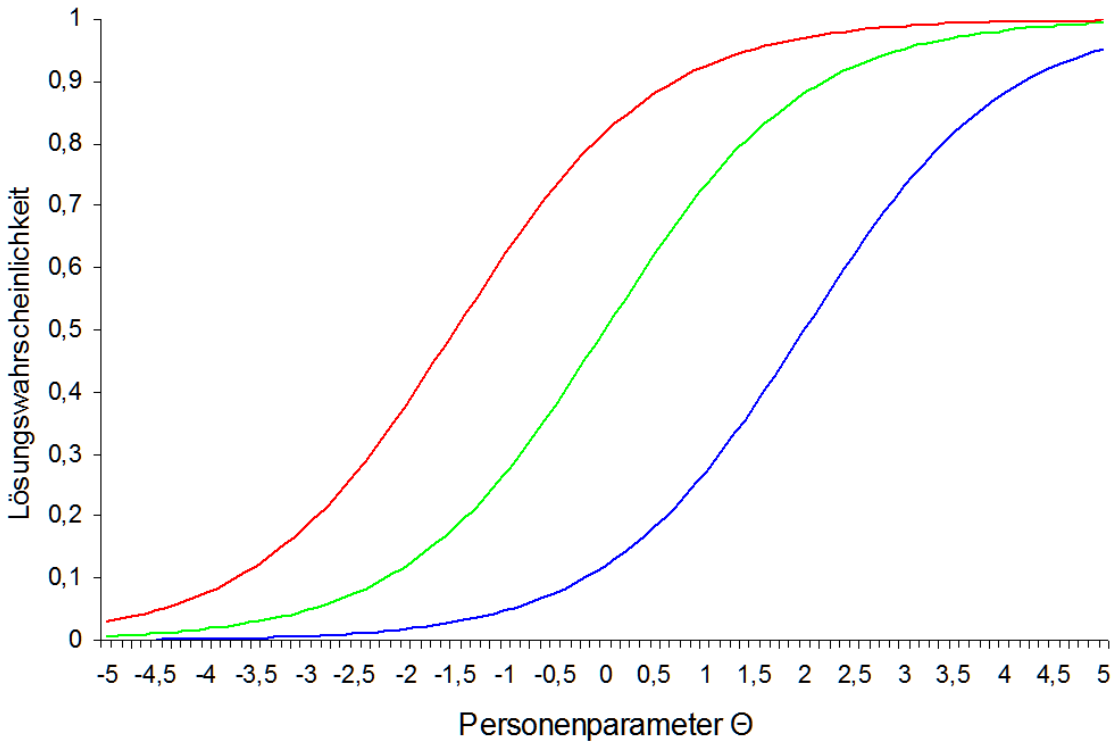


Abbildung 24 Zusammenhang zwischen dem Personenparameter und der Lösungswahrscheinlichkeit im Rasch Modell für drei Aufgaben mit unterschiedlichen Schwierigkeiten

Wie in Abbildung 24 dargestellt, steigt die Wahrscheinlichkeit, eine Aufgabe richtig zu lösen, mit zunehmender Fähigkeit einer Person. Die drei dargestellten Funktionen entsprechen drei Aufgaben, die sich in ihrer Schwierigkeit (β_j) unterscheiden, die auch als Lage des Funktionsgraphen relativ zur X-Achse verstanden werden kann. Hat eine Person einen Personenparameter von $\Theta_i = 0$, so beträgt ihre Lösungswahrscheinlichkeit für die rot dargestellte Aufgabe 81,76 %, die Wahrscheinlichkeit, die grün dargestellte Aufgabe zu lösen, beträgt 50 % und die Wahrscheinlichkeit, die blau dargestellte Aufgabe zu lösen, 11,92 %. Als Item- oder Aufgabenschwierigkeit wird der Wert auf der X-Achse bezeichnet, an dem die Lösungswahrscheinlichkeit einer Aufgabe 50 % beträgt; diese Stelle entspricht dem Wendepunkt der jeweiligen Kurve. Für die rot dargestellte Aufgabe beträgt die Aufgabenschwierigkeit $\beta = 1{,}5$, für die grün dargestellte Aufgabe $\beta = 0$ und für die blau dargestellte Aufgabe $\beta = 2$.

Das klassische dichotome Rasch-Modell kann durch seine Beschränkung auf die zwei Lösungsalternativen, eine Aufgabe richtig oder falsch zu lösen, den Fall, dass eine Aufgabe teilweise richtig gelöst wurde, nicht abbilden. Auch für die in der psychologischen Diagnostik gängigen polytomen Ratingskalen und für die Verarbeitung von metrischen Daten, die durch das Auszählen von Häufigkeiten entstanden sind, kann das dichotome Rasch Modell nicht genutzt werden. Deshalb wurde von Rasch (1961) ein Ansatz formuliert, um auch polytome Antwortformate berücksichtigen zu können. Darauf aufbauend wurde von Masters (1982) das Partial Credit Modell (PCM) formuliert, dessen Modellgleichung Formel 2 darstellt:

$$P(U_{ij}) = c \mid \theta_i, \beta_j) = \frac{e^{c*\theta_i - \beta_{jc}}}{\sum_{l=0}^{m_j} e^{l*\theta_i - \beta_{jl}}} \tag{2}$$

Im PCM weist jede Aufgabe j i = 0, … m_j Antwortkategorien auf. Die Gleichung 2 gibt die Wahrscheinlichkeit an, dass eine bestimmte Antwortkategorie c gewählt wird, in Abhängigkeit von der Merkmalsausprägung der Person (Θ_i) und der Schwierigkeit dieser Antwortkategorie (β_{jc}). Im PCM ist es analog zum Rasch-Modell möglich, den Zusammenhang zwischen der Fähigkeit einer Person und der Wahrscheinlichkeit, eine bestimmte Antwortkategorie auszuwählen, grafisch darzustellen. Abbildung 25 zeigt diesem Zusammenhang exemplarisch für Item 6 dieser Studie.

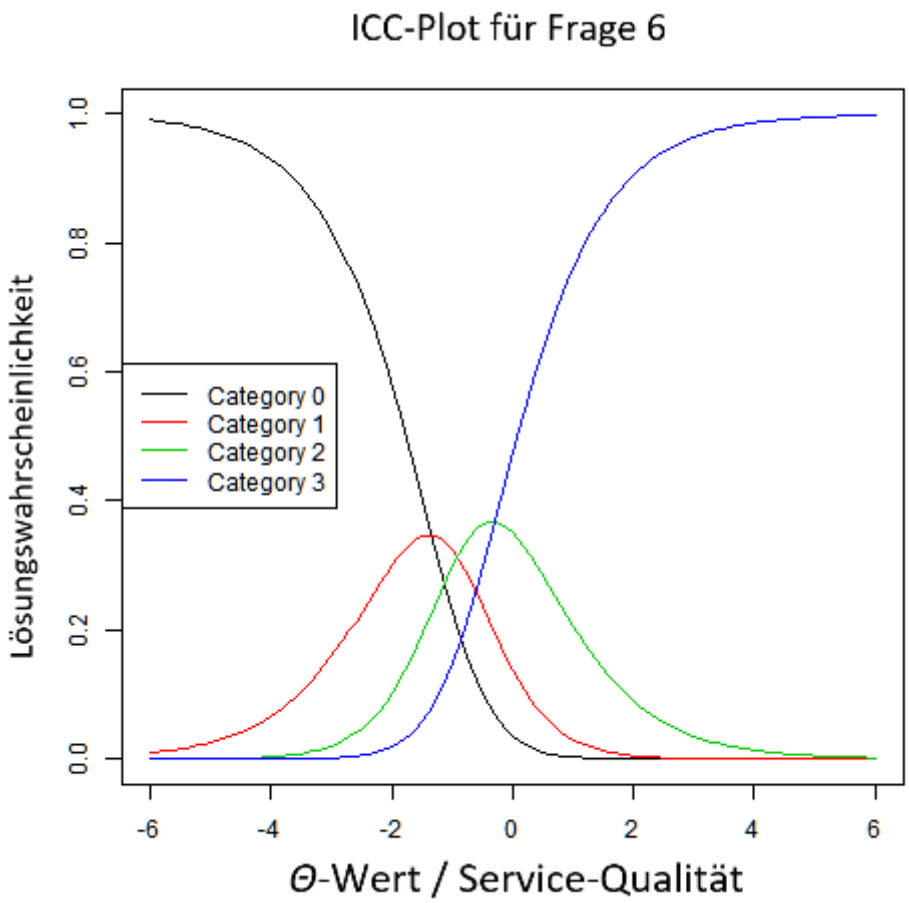


Abbildung 25 Zusammenhang zwischen dem Personenparameter (Service-Qualität) und der Wahrscheinlichkeit, dass eine Antwortkategorie (Category 0 – 3) gewählt wird am Beispiel von Item 6 dieser Studie

Das PCM ermöglicht es, dass Items, die zu einem latenten Merkmal zusammengefasst werden, unterschiedlich viele Antwortkategorien haben. Darüber hinaus kann der Abstand zwischen den Antwortkategorien auf der latenten Dimension von Item zum Item unterschiedlich groß sein. Zudem macht das PCM keine Aussagen darüber, ob die einzelnen Antwortkategorien in einer auf- oder absteigenden Reihenfolge entlang der latenten Dimension angeordnet sind. Für Item 6 (Abbildung 25) dieser Studie würde ein PCM die Parameter der Antwortkategorien so schätzen, dass Personen mit einer sehr geringen Ausprägung (Θ < -2) auf der zugrunde liegenden latenten Dimension mit hoher Wahrscheinlichkeit die Antwortkategorie 0 (trifft nicht zu) wählen würden (schwarze Kurve in Abbildung 25). Personen mit einem Personenparameter von Θ = -1,6 würden sich am wahrscheinlichsten für die Antwortkategorie 1 (rote Kurve in Abbildung 25) entscheiden. Mit weiter ansteigendem Personenparameter ist die Wahrscheinlichkeit für Antwortkategorie 2 (grüne Kurve in Abbildung 25) und ab einem Parameter von Θ = -0,2 die Wahrscheinlichkeit für Antwortkategorie 3 „trifft voll zu“ (blaue Kurve in Abbildung 25)

am höchsten. Am Beispiel von Item 6 kann man mittels PCM zeigen, dass die vier Antwortkategorien wie intendiert in einer aufsteigenden Abfolge entlang des Personenparameters angeordnet sind. Im Anhang E sind die Übergänge von einer Antwortkategorie zur nächsten als Schwellen für alle Items dieser Studie dargestellt. Items, bei denen diese Schwellen zwischen den Antwortkategorien nicht aufsteigend entlang des Personenparameters angeordnet sind, werden rot hervorgehoben. Gorin, Dodd, Fitzpatrick und Shieh (2005) zeigen, wie die Item-Parameter aus einem PCM-Modell für computerbasierte adaptive Tests genutzt werden können.

Um das Antwortverhalten von Versuchspersonen bei klassischen Fragebogenstudien, in denen alle Items die gleiche Anzahl von Antwortkategorie nutzen, im Rahmen der probabilistischen Testtheorie modellieren zu können, hat Andrich (1978) das sogenannte Rating Scale Model (RSM) vorgeschlagen. Das RSM wurde zeitlich vor dem PCM formuliert und bezieht sich direkt auf den Ansatz von Rasch (1961), der polytome Antwortformate berücksichtigt.

$$\ln\left(\frac{P_{nij}}{P_{nij-1}}\right) = \beta_n - (\delta_i + \tau_j) \tag{3}$$

Beim RSM, das in Gleichung 4 formuliert ist, wird die Wahrscheinlichkeit, dass einen Person n genau j Schwellenwerte auf dem Item i überschreitet, mit P_{nij} bezeichnet. P_{nij-1} bezeichnet analog dazu die Wahrscheinlichkeit, dass eine Person genau j-1 Schwellen überschreitet. β_n bezeichnet die Trait-Ausprägung einer Person n und δ_i die Lage eines Items i auf der latenten Dimension. τ_j bezeichnet den Schwellwert-Parameter, der die Grenze zwischen zwei Antwortkategorien relativ zur Lage des Items entlang der latenten Dimension angibt. Beim RSM handelt es sich mathematisch betrachtet um eine Vereinfachung des PCM, die darin besteht, dass alle Items die gleiche Anzahl von Antwortkategorien aufweisen müssen und darüber hinaus die Abstände zwischen den Antwortkategorien für alle Items gleichgesetzt werden (Sundström, 2011). Der Parameter τ_j wird nur einmal für alle Items geschätzt. Zudem wird im RSM im Kontrast zum PCM definiert, dass die einzelnen Antwortkategorien eines Items in einer Rangreihe stehen (Ostini & Nering, 2006). Im Anhang F werden die Übergänge von einer Antwortkategorie zu nächsten als Schwellen gemäß dem RSM für alle Items dieser Studie dargestellt. Im Vergleich zu Anhang E sieht man hier, dass im RSM die geschätzten Abstände zwischen den Antwortkategorien für alle Items gleich sind.

Um in dieser Studie die Item-Parameter für die Items der Skala Service-Qualität zu bestimmen, wurde auf ein RSM zurückgegriffen, weil die Annahmen dieses Modells sehr gut auf die intendierte Anwendung des Antwortformates passen. Alle Items nutzten das gleiche Antwortformat und das Antwortformat soll für alle Items gleichermaßen funktionieren. Mit dem RSM gelingt darüber hinaus eine stabilere Schätzung der Item-Parameter als mit dem PCM. Wie in Abbildung 17 deutlich wird, machten die Versuchspersonen bei einigen Items nur sehr selten

von der Antwortkategorie „trifft nicht zu“ Gebrauch. Da im PCM für jede Antwortkategorie aller Items ein eigener Parameter geschätzt wird, würde diese Schätzung bei einigen Items für die Antwortkategorien „trifft nicht zu“ auf einer sehr geringen Anzahl von Beobachtungen basieren, wodurch die Schätzung mit einer relativ großen Unsicherheit behaftet ist.

7.4.2 Wahl des ersten Items

Welches Item in einem adaptiven Test als erstes Item dargeboten wird, kann grundsätzlich statisch festgelegt werden. Liegen bereits Informationen über die Ausprägung der Versuchsperson auf dem zu messenden Merkmal vor, kann dies bei der Wahl des ersten Items berücksichtigt werden. So könnte zum Beispiel bei der Diagnostik von Hochbegabung bereits mit einem Item begonnen werden, das eine höhere Item-Schwierigkeit aufweist. Wird ein adaptives Testverfahren in die Erfassung weiterer Merkmal eingebettet, die es ermöglichen, erste Prognosen über die Ausprägung der Person auf dem adaptiv zu bestimmenden Konstrukt zu treffen, so ist es sinnvoll, im Sinne einer Verkürzung des Tests und der damit gesteigerten Ökonomie diese erste Prognose als Ausgangspunkt für die Bestimmung des ersten Test-Items zu nutzen.

Wie in Kapitel 7.3.4 beschrieben, kann die Abfolge der Items Einfluss auf das Messergebnis nehmen. Bedenkt man den beschriebenen „warm-up“-Effekt des ersten Items, der besagt, dass dieses dazu dient, die Versuchsperson mit dem Antwortformat vertraut zu machen, so sollte man das erste Item für alle Teilnehmenden konstant halten (Walter & Rose, 2013, S. 89).

Der in dieser Arbeit entwickelte adaptive Test wird zunächst gesondert von anderen Testverfahren dargeboten und kann daher bei der Wahl des ersten Items auf keine vorliegenden Informationen zurückgreifen. Aus diesem Grund und um den Beginn des Tests für alle Teilnehmenden möglichst einheitlich zu gestalten, wurde das erste Test-Item statisch definiert. Bei der Wahl des ersten Items wurde darauf geachtet, ein Item zu nutzen, das eine mittlere Item-Schwierigkeit hat, denn es wird zunächst von der Annahme ausgegangen, dass eine Versuchsperson bzw. eine Organisation eine mittlere Ausprägung auf dem Konstrukt Service-Qualität aufweist. Diese Annahme ist plausibel, wenn keine weiteren Informationen vorliegen, und führt über die Testung vieler Personen bzw. Organisationen hinweg dazu, dass durchschnittlich die beste a priori-Prognose der Merkmalsausprägung getroffen wird.

7.4.3 Schätzung des Personenparameters und dessen Standardfehlers

Um den Personenparameter Θ anhand der vorliegenden Antworten zu schätzen, haben sich zwei Verfahren etabliert, die zum einen auf der Maximum-Likelihood-Methode und zum anderen auf einem Bayes-Schätzer basieren. Beide Verfahren liefern einen Punktschätzer für Θ und ermöglichen es darüber hinaus, einen Standardfehler zu bestimmen, anhand dessen ein

Konfidenzintervall um den Punktschätzer gelegt werden kann. Die Schätzung des Personenparameters Θ anhand der Maximum-Likelihood (ML) -Methode basiert auf der Likelihood-Funktion, die in Formel 4 für ein dichotomes Rasch Modell dargestellt wird (Baker, 1992, S. 66).

$$P(U_j \mid \theta) = \prod_{i=1}^{n} P_i^{u_{ij}}(\theta_j) Q_i^{1-u_{ij}}(\theta_j) \tag{4}$$

Die Likelihood-Funktion beschreibt die Wahrscheinlichkeit des Antwortvektors U_j, der alle Antworten einer Versuchsperson j auf die Items i des Tests enthält, gegeben eines Personenparameters Θ. Ziel der Θ-Schätzung mit diesem Verfahren ist es, einen Wert für Θ zu finden, für den das beobachtete Antwortmuster die höchstmögliche Wahrscheinlichkeit hat. Löst man Gleichung 4 nach Θ_j auf und setzt das Newton-Raphson-Verfahren ein, um das Maximum der Likelihood-Funktion iterativ zu bestimmen, kann eine Schätzung des Personenparameters ($\widehat{\theta_j}$) durchgeführt werden (Embretson & Reise, 2000; Kuk & Cheng, 1997). Der Standardfehler des $\widehat{\theta_j}$-Wertes lässt sich wie in Gleichung 5 beschrieben bestimmen.

$$SE_{\widehat{\theta}_j} = \sqrt{s^2_{\widehat{\theta}_j}} \tag{5}$$

Das zweite etablierte Schätzverfahren für den Personenparameter Θ, das in der Literatur als EAP-Schätzung (expected a posteriori) bezeichnet wird, basiert auf dem Satz von Bayes, der sich mit bedingten Wahrscheinlichkeiten befasst (van der Linden, Wim J. & Ren, 2020). Gleichung 6 stellt das Bayes-Theorem angewandt auf die Schätzung von Θ dar.

$$P(\theta \mid u) = \frac{P(\theta \mid u)\, k(\theta)}{P(u)} \tag{6}$$

Die Wahrscheinlichkeit eines bestimmten Personenparameters gegeben eines bestimmten Antwortvektors $P(\Theta \mid u)$ wird als Quotient aus dem Produkt dieser bedingten Wahrscheinlichkeit mit der angenommenen Verteilung des Parameters k (Θ) und der der Wahrscheinlichkeit $P(u)$, mit der ein Antwortvektor u auftritt, beschrieben (Bock & Mislevy, 1982).

Vergleicht man die beiden etablierten Verfahren zur Bestimmung des Personenparameters Θ, so fällt zunächst auf, dass ein Nachteil der ML-Methode darin besteht, dass sie zu keiner eindeutigen Θ-Schätzung kommt, wenn eine Versuchsperson maximales bzw. minimales Antwortverhalten aufweist. Beantwortet eine Versuchsperson zum Beispiel alle Items durch Wahl der höchste Antwortkategorie („trifft voll zu"), kann für diese mit der ML-Methode kein Θ-Wert geschätzt werden (Keller, 2000). Obwohl dieses Szenario sehr unwahrscheinlich ist, da in der Regel bereits bei der Testkonstruktion darauf geachtet wurde, Items mit sehr hoher und sehr

niedriger Item-Schwierigkeit in den Item-Pool zu integrieren, ist es doch denkbar. Insbesondere für adaptive Tests ist diese Eigenschaft der ML-Schätzung kritisch, da nach jedem Item eine Schätzung des Personenparameters durchgeführt wird. Gerade zu Beginn eines adaptiven Tests ist die Wahrscheinlichkeit relativ hoch, dass eine Versuchsperson alle bislang dargebotenen Items maximal bzw. minimal beantwortet hat und deshalb keine Schätzung von Θ möglich ist. Dies ist problematisch, weil die Schätzung des Personenparameters, wie in Abbildung 23 dargestellt, die Grundlage für die Auswahl des nächsten Items ist. Um dieses Problem zu lösen, kann der Θ-Wert bei extremem Antwortverhalten auf extreme Werte von z. B. –4 bzw. +4 fixiert werden (Han, 2016). Hambleton, Swaminathan und Rogers (1995) schlagen für diesen Fall eine modifizierte ML-Methode vor, bei der in solchen extremen Fällen ein kleiner Wert zum aktuellen Schätzer addiert (bei maximalem Antwortverhalten) bzw. subtrahiert (bei minimalem Antwortverhalten) wird. Ein weiteres Problem der ML-Methode kann entstehen, wenn die Likelihood-Funktion mehrere Maxima aufweist und die genutzten Startwerte für das iterative Schätzverfahren lediglich ein lokales Maximum identifizieren (Myung, 2003). Die wiederholte Anwendung des Verfahrens mit unterschiedlichen Startwerten reduziert diese Problematik, führt jedoch dazu, dass das ML-Verfahren rechenaufwendiger und verglichen mit einer EAP-Schätzung zeitaufwendiger ist (van der Linden, Wim J. & Ren, 2020).

7.4.4 Kriterien für die Beendigung eines adaptiven Tests

Die Beendigung eines adaptiven Tests kann von verschiedenen Kriterien abhängig gemacht werden. Abbildung 26 zeigt die in dieser Arbeit genutzte Abbruchlogik, die den Standardfehler der Messung und eine Mindest- und eine Maximalanzahl von Items umfasst. Es wird geprüft, ob der Standardfehler der aktuellen Θ-Schätzung einen vorab definierten Wert unterschreitet. Ist dieses Kriterium erfüllt, wird anschließend geprüft, ob die Mindestanzahl von vorgegebenen Items erreicht oder überschritten ist. Ist auch dieses Kriterium erfüllt, wird der Test beendet. Ist der Standardfehler der aktuellen Theta-Schätzung größer als der vorab definierte Wert, wird geprüft, ob die Anzahl der bislang dargebotenen Items die definierte maximale Anzahl von Items erreicht hat. Ist dieses Kriterium erfüllt, so wird der Test beendet, andernfalls wird er fortgesetzt.

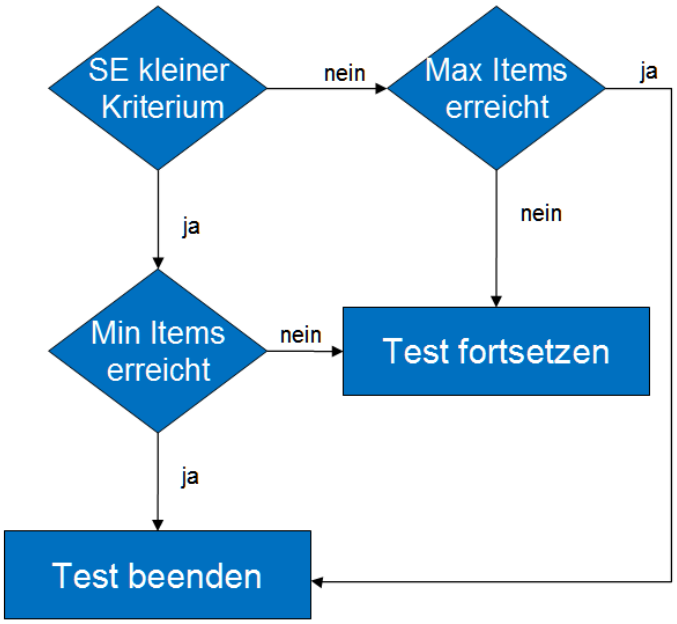


Abbildung 26 Schematische Darstellung der gewählten Abbruchlogik; SE = Standardfehler der Messung

Die Wahl des zu unterschreitenden Standardfehlers der Messung orientierte sich an der Reliabilität konventioneller Testverfahren. Die klassische Testtheorie beschreibt den Standardfehler der Messung (SE_M), wie in Formel 7 dargestellt, als Produkt der Standardabweichung des Tests (σ) mit der Quadratwurzel von 1 minus der Reliabilität (r_{tt}) (Embretson & Reise, 2000, S. 16).

$$SE_m = (1 - r_{tt})^{\frac{1}{2}} \sigma \tag{7}$$

Will man im Rahmen eines adaptiven Tests aus dem durchschnittlichen Standardfehler ($\overline{SE}$) auf die Reliabiltät des Tests schließen, so kann man Gleichung 7 umstellen und erhält die folgende Gleichung 8.

$$r_{tt} = 1 - \frac{\overline{SE}^2}{\sigma^2} \tag{8}$$

Da im Rahmen der adaptiven Testung der Personenparameter Θ so bestimmen wird, dass er normalverteilt mit einem Mittelwert von 0 und einer Standardabweichung von 1 ist, kann für σ^2 der Wert 1 angenommen werden. Damit vereinfacht sich Gleichung 8 zu:

$$r_{tt} = 1 - \overline{SE}^2 \tag{9}$$

Embretson und Reise (2000) zeigen anhand simulierter Daten, dass sich die Reliabilität der Messung zwischen linearen und adaptiven Tests grundsätzlich unterscheidet. Wie in Abbildung 27 dargestellt, steigt der Standardfehler der Messung in den Extrembereichen der Trait-Ausprägung bzw. der latenten Dimension bei klassischen, linearen Testverfahren parabelförmig an. Dementsprechend ist die Reliabilität der Messung bei diesen Verfahren für seltene Fälle mit extrem hohen oder niedrigen Ausprägungen deutlich geringer. Verglichen damit liefern adaptive Testverfahren über den gesamten Bereich der latenten Dimension Messergebnisse mit gleichmäßig hoher Reliabilität bzw. geringem Standardfehler. Die Ergebnisse der Simulationsstudie von Embretson und Reise (2000) setzen voraus, dass die Item-Bank, die dem adaptiven Test zugrunde liegt, den gesamten Bereich der Item-Schwierigkeit suffizient abdeckt. Ist dies gegeben, kann angenommen werden, dass adaptive Tests auch in den Bereichen extremer Θ-Ausprägung genaue Schätzungen der Trait-Ausprägung ermöglichen. Die Studie von Embretson und Reise (2000) zeigt darüber hinaus, dass Verfahren, die eine höhere Anzahl von Test-Item s nutzen, eine höhere Reliabilität aufweisen. Für adaptive Verfahren bedeutet dies, dass die Reliabilität der Messung normalerweise durch die Nutzung weiterer Items gesteigert werden kann.

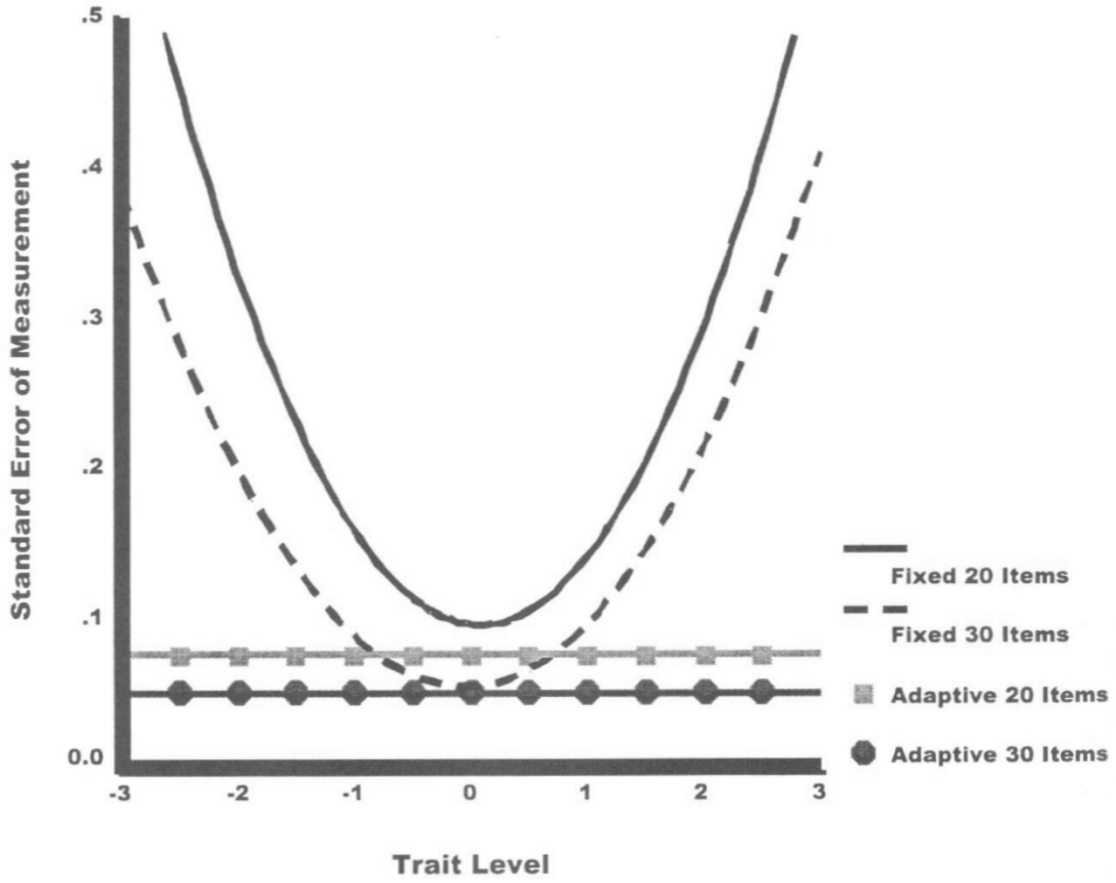


Abbildung 27 Zusammenhang zwischen Trait-Ausprägung und dem Standardfehler der Messung für unterschiedliche Testlängen und Testformen (Embretson & Reise, 2000)

Diese Ausführungen machen deutlich, dass man bei der Konstruktion eines adaptiven Tests durch die Wahl eines Abbruchkriteriums, das sich am Standardfehler der Messung orientiert, direkt Einfluss auf die Reliabilität der Messung nehmen kann. Die Messgenauigkeit einer jeden einzelnen Messung kann beurteilt werden, weil für jede einzelne Θ-Schätzung ein Standardfehler angegeben werden kann. Die Reliabilität der Messung kann für jeden Einzelfall beurteilt werden, was verglichen mit der klassischen Testtheorie, bei der üblicherweise nur ein Reliabilitätsschätzer für einen gesamten Test geschätzt wird, ein Vorteil ist. Soll anhand des Ergebnisses des Tests eine binäre Entscheidung getroffen werden, kann der Standardfehler der Messung genutzt werden, um ein Konfidenzintervall um den Punktschätzer zu legen.

Bei einigen adaptiven Tests wird als Abbruchkriterium $SE <= 0.32$ eingesetzt (Fliege et al., 2005; Kocalevent, 2005; Walter et al., 2007). Geht man davon aus, dass mit einem solchen Abbruchkriterium eine Θ-Schätzung gelingt, die einen durchschnittlichen Standardfehler von $\overline{SE} = 0.32$ aufweist, ergibt sich für diese Konfiguration gemäß Gleichung 9 eine Reliabilität von $r_{tt} = .90$.

Für Personen, die den Test vorzeitig abbrechen, bleibt die Θ-Schätzung mit einem relativ hohen Standardfehler verknüpft. Aus diesen Gründen führt ein Abbruchkriterium von $SE = 0.32$ in der Praxis zu einem durchschnittlichen Standardfehler ($\overline{SE}$), der etwas höher liegt, so dass die tatsächliche durchschnittliche Reliabilität des Tests geringer ausfällt als beabsichtigt. Um dieses Phänomen zu kompensieren, wurde in der Simulationsstudie ein etwas konservativeres Abbruchkriterium $SE <= 0.30$ gewählt. Würde es mit diesem Kriterium gelingen, einen durchschnittlichen Standardfehler der Messung von $\overline{SE} = 0.30$ zu erzielen, so würde dies rechnerisch gemäß Gleichung 9 einer Reliabilität von $r_{tt} = .91$ entsprechen. Üblicherweise wird die Reliabilität eines psychodiagnostischen Messverfahrens ab einem Wert von > .70 als akzeptabel angesehen. Da der in dieser Arbeit entwickelte adaptive Test primär zur Einzelfalldiagnostik eingesetzt werden soll, erscheint eine solche hohe Reliabilität der Messung als Zielgröße als angemessen.

In das Abbruchkriterium eines adaptiven Tests können neben den bisher beschriebenen noch weitere Variablen einfließen. Choi, Grady und Dodd (2010) schlagen vor, die vorhergesagte Reduktion des Standardfehlers der Messung, die bei der Vorgabe eines weiteren Items erwartet wird, zu berücksichtigen. Eine weitere Variable, die beim Abbruchkriterium beachtet werden kann, ist die Bearbeitungsdauer. Es kann eine minimale bzw. maximale Bearbeitungszeit festgelegt werden, die in das Abbruchkriterium integriert wird (Wainer & Dorans, 2000). Wenn sichergestellt werden soll, dass aus jedem Bereich des Konstrukts eine Mindestanzahl von Items dargeboten wird, muss dies im Rahmen eines „Content Balancing“ ebenfalls bei der Bildung des Abbruchkriteriums berücksichtigt werden (Cheng, Chang & Yi, 2007).

7.4.5 Algorithmen zur Auswahl des nächsten Test-Items

Nicht adaptive Verfahren umfassen meistens eine feste Anzahl und Abfolge von Items. Bei der Wahl des nächsten Items folgen sie einer sehr einfachen Strategie und bieten alle verfügbaren Items linear in einer festgelegten Reihenfolge dar. Adaptive Tests zeichnen sich dadurch aus, dass die Wahl des nächsten Items abhängig von den ermittelten Item-Parametern und dem geschätzten Personenparameter erfolgt. Je nach Anwendungsbereich und Item-Format gibt es unterschiedliche Algorithmen, die bei der Wahl des nächsten Items genutzt werden können.

Werden adaptive Tests eingesetzt, um das Vorhandensein oder Fehlen spezifischer Fertigkeiten im Zusammenhang mit pädagogischen Bewertungen eingesetzt, so wird dies in der englischen Fachliteratur als „cognitive diagnosis" bezeichnet. In sogenannten „cognitive diagnotic models" wird die Wahrscheinlichkeit, ein Item richtig zu beantworten, als Funktion eines Musters an bestimmten Fähigkeiten beschrieben (Henson, Roussos, Douglas & He, 2008). Im Gegensatz zur Item-Response-Theorie wird nicht eine latente Dimension zur Vorhersage der Lösungswahrscheinlichkeit herangezogen, sondern latente Klassen angenommen, die einem bestimmten Profil an Fähigkeiten entsprechen und vorhersagen sollen, wie wahrscheinlich ein Item richtig gelöst wird. Da diese Art von Messmodellen sich grundlegend von IRT-basierten Modellen unterscheiden, werden in diesem Anwendungsbereich spezielle Algorithmen zur Bestimmung des nächsten Items benutzt (Kaplan, La Torre & Barrada, 2015; Wang, 2013; Zheng & Chang, 2016).

Traditionell basieren adaptive Tests auf der probabilistischen Testtheorie und den Modellen der Items-Response-Theorie (IRT), die in Kapitel 7.4.1 beschrieben wurden. Diese Modelle sind in der Regel eindimensional, das heißt alle Items dienen dazu, Informationen über eine zugrunde liegende latente Dimension zu erhalten. Die Item-Parameter und der Personenparameter werden auf einer gemeinsamen Skala, die dieser latenten Dimension entspricht, abgebildet. Für die Wahl eines nächsten Items ist vor allem die Item-Schwierigkeit relevant, die die Lage eines Items entlang dieser Skala beschreibt. In Anhang F ist die Lage aller Items der Skala Service-Qualität relativ zur latenten Dimension dargestellt. Bei den meisten Selektionsstrategien wird ausgehend von dem geschätzten Personenparameter dasjenige Item gesucht, das für diesen vorläufig angenommenen Personenparameter am meisten Information liefert. Die Information, die die Beantwortung eines weiteren Items liefert, kann als Funktion der Item-Schwierigkeit und dem Diskriminationsparameter bzw. bei polytomen Antwortformaten der Lage der Schwellwerte des Items beschrieben werden (Muraki, 1993). Die bekannteste Methode, die auf dieser Item-Informations-Funktion basiert, wird als „maximum Fischer information"-Algorithmus (MFI) bezeichnet (Barrada, Olea, Ponsoda & Abad, 2009). Insbesondere für adaptive Tests, die auf Items basieren, die ein polytomes Antwortformat nutzen, wurden von Choi und Swartz (2009) neben dem MFI-Algorithmus mit Maximum Likelihood Weighted Information (WLWI), Maximum Posterior Weighted Information (MPWI), Maximum Expected Information (MEI),

Minimum Expected Posterior Variance (MEPV) und Maximum Expected Posterior Weighted Information (MEPWI) fünf weitere Selektionsstrategien vorgestellt. Um diese Selektionsstrategien vergleichen zu können, entwickelten und nutzten sie das das Tool Firestar (Choi, 2009; Choi, Podrabsky & McKinney, 2012). Anhand empirischer und simulierter Datensätzen verglichen Choi und Swartz (2009) Verlauf und Testergebnisse für die verschiedenen Selektionsstrategien. Sie stellten fest, dass alle beschriebenen Verfahren zur Selektion des nächsten Test-Items ähnlich abschnitten. Obwohl theoretisch zu erwarten war, dass die komplexeren Selektionsalgorithmen überlegen sind, erwies sich die MFI-Strategie in Verbindung mit einer Θ-Schätzung nach dem bayesianischen EAP-Verfahren als konkurrenzfähig. Für Item-Pools mit einer geringen Anzahl von polytomen Items erschienen alle beschriebenen Methoden angemessen. Welches Verfahren zur Bestimmung des nächsten Test-Items für einen spezifischen Item-Pool überlegen ist, lässt sich nicht pauschal sagen, weshalb bei der Wahl des Selektionsalgorithmus empfohlen wird, Simulationsstudien zu nutzen (Choi et al., 2012).

Wird die Wahl des nächsten Items ausschließlich von rein psychometrischen Aspekten abhängig gemacht, kann es vorkommen, dass einzelne Personen nur sehr spezifische Items dargeboten bekommen. Wenn zum Beispiel bestimmte Persönlichkeitseigenschaften oder psychische Störungen diagnostiziert werden sollen, kann eine solche Item-Selektion dazu führen, dass das Konstrukt nicht mehr in seiner theoretisch begründeten vollen Breite erfasst wird. Theoretisch und inhaltlich relevante Items können durch die Item-Auswahl entlang psychometrischer Kriterien ausgelassen werden. Aus diagnostischer Sicht kann es relevant sein, dass zumindest ein Teil der Items bei allen Versuchspersonen gleichermaßen erfasst wird. Deshalb können bei der Item-Selektion zusätzlich sogenannte Testskripts berücksichtigt werden, die dafür sorgen, dass die Testinhalte ausbalanciert werden und sichergestellt wird, dass alle Versuchspersonen eine Mindestanzahl von Items aus bestimmten Bereichen des Konstrukts dargeboten bekommen (Cheng et al., 2007; Han, 2018).

Im Bereich der Leistungsdiagnostik wird in der Regel gefordert, dass die Items der eingesetzten Testverfahren für die Versuchspersonen geheim sind. Dies ist erforderlich, damit der Test für alle teilnehmenden Personen gleich schwierig ist und Fairness und Validität des Verfahrens gewahrt bleiben. Werden durch die Item-Selektion in adaptiven Tests einzelne Items besonders häufig ausgewählt, kann dies dazu führen, dass die Inhalte und Lösungen dieser Items durch Personen, die am Testverfahren teilgenommen haben, weitergegeben werden und einzelnen künftigen Testpersonen bereits bekannt sind. Um dies zu vermeiden, wurden verschiedene Strategien entwickelt, um die Exposition der Items zu kontrollieren (Chang & Ansley, 2003; Öztürk & Dogan, 2015). Bei der Anwendung adaptiver Tests, z. B. in medizinischen Diagnosefragebögen, Persönlichkeitsmessungen oder adaptiven Lernmitteln, ist die Kontrolle der Exposition der Items nicht notwendig ist (Han, 2018).

7.5 Simulationsstudie

Um einen webbasierten adaptiven Test entwickeln und anbieten zu können, müssen die in Kapitel 7.6 beschriebenen notwendigen technischen Voraussetzungen in Form eines entsprechend ausgestatteten Webservers geschaffen werden. Zudem muss der in Kapitel 7.4 beschriebene Ablauf eines adaptiven Tests als dynamische Website programmiert werden. Um den Inhalt der Website dynamisch generieren zu können, wurden die Open Source-Skriptsprachen PHP und das Open Source-Statistikprogramm R genutzt. Im Rahmen der Realisierung des webbasierten adaptiven Tests für Service-Qualität wurden HTML-Seiten sowie PHP- und R-Skripte konzipiert, programmiert, getestet und integriert. Wie diese Skripte integriert wurden, wird in Kapitel 7.7 zusammengefasst. Bei der Programmierung der PHP- und R-Skripte, die die grundlegenden Funktionen des adaptiven Tests abbilden, mussten verschiedene Entscheidungen getroffen werden. So musste festgelegt werden, nach welchem Algorithmus das erste Fragebogen-Item und die weiteren Items ausgewählt werden, wie anhand des Antwortvektors der Personenparameter und dessen Standardfehler geschätzt werden und nach welchen Kriterien das Ende des Tests definiert wird. Um diese Entscheidungen zu treffen, konnte nur bedingt auf die Ergebnisse der bereits vorliegenden Studien zurückgegriffen werden, da diese Entscheidungen unter anderem von den zur Verfügung stehenden Items und deren psychometrischen Messeigenschaften abhängig sind. Zudem musste festgelegt werden, auf welches Ziel hin der adaptive Test optimiert werden soll. Ein Ziel könnte sein, die Ökonomie des Verfahrens zu optimieren, indem mit möglichst wenigen Fragen eine akzeptabel genaue Schätzung des Personenparameters erreicht wird. Ein anderes Ziel könnte sein, möglichst exakte Schätzungen des Personenparameters zu erhalten und dafür gegebenenfalls eine höhere Anzahl von Test-Items zu nutzen. Die Algorithmen und Verfahren zur Schätzung des Personenparameters und dessen Standardschätzfehler und zur Auswahl des nächsten Test-Items, die in Kapitel 7.4 dargestellt wurden, sind unterschiedlich rechen- und damit ressourcen- und zeitaufwendig und führen zu leicht unterschiedlichen Ergebnissen. Wenn ein System entwickelt werden soll, bei dem möglichst viele Personen gleichzeitig getestet werden können, hat eine hohe Performanz der gewählten Lösungen hohe Priorität. In diesem Fall können diese Entscheidungen zu Gunsten einer möglichst wenig rechenintensiven Lösung ausfallen, die trotzdem akzeptabel genaue Ergebnisse liefert.

Da je nach Optimierungsziel und vorliegendem Item-Pool unterschiedliche Konfigurationen optimale Ergebnisse liefern, wurde das in dieser Arbeit entwickelte System so konzipiert, dass es ermöglicht, über die Anpassung der Konfiguration zwischen verschiedenen Einstellungen hin- und herzuwechseln. Diese Konzeption ermöglicht eine experimentelle Vorgehensweise, um die verschiedenen Konfigurationen im Hinblick auf das gewählte Optimierungsziel zu vergleichen. Die Versuchspersonen können im Sinne eines klassischen Versuchsplans zufällig verschiedenen Konfigurationen bzw. Versuchsbedingungen zugeordnet werden. Würde man die so erfassten Daten mit einem gleichzeitig erfassten Validitätskriterium in Verbindung bringen, könnte experimentell

geprüft werden, welche Konfigurationen sich vor dem Hintergrund spezifischer Optimierungsziele als besonders sinnvoll erweisen. Diese Herangehensweise wäre beim Vergleich vieler verschiedener Konfigurationen sehr aufwendig und würde eine hohe Zahl von Versuchspersonen benötigen.

Alternativ zu dieser zeit- und kostenintensiven experimentellen Herangehensweise besteht die Möglichkeit, im Rahmen einer Simulationsstudie die verschiedenen Konfigurationen anhand der bereits vorliegenden Daten zu testen. Dieser Ansatz erfordert keine Erfassung neuer Datensätze und stellt eine sehr ökonomische Möglichkeit dar, um zu prüfen, wie die Testung und das Testergebnis für die bereits erfassten Datensätze bei bestimmten Konfigurationen des adaptiven Tests ausgefallen wären. Der Vergleich der Ergebnisse eines adaptiven Tests mit einer bestimmten Konfiguration mit den Testergebnissen der klassischen nicht adaptiven Testung, die bereits mit dem ersten Teil dieser Arbeit vorliegt, ist eine Möglichkeit, die Validität der gewählten Konfiguration nachzuweisen.

7.5.1 Fragestellung der Simulationsstudie

Ziel der durchgeführten Simulationsstudie war es herauszufinden, unter welchen Bedingungen die Ergebnisse eines adaptiven Tests mit einer bestimmten Konfiguration mit den Testergebnissen einer klassischen linearen Testung mittels aller Test-Items der Skala zur Erfassung von Service-Qualität vergleichbar sind. Um dies zu beurteilen, wurden die Zusammenhänge der Testergebnisse der adaptiven Testung mit den Testergebnissen der klassischen Testung und der Theta-Schätzung anhand des vollständigen Antwortvektors mittels Korrelationskoeffizienten verglichen. Da ein häufiges Ziel bei der Entwicklung adaptiver Tests die Verkürzung der Testlänge ist, wurde darüber hinaus festgehalten, wie viele Items pro Versuchsperson durchschnittlich zur ausreichend präzisen Schätzung des Personenparameters benötigt wurden.

Wie in Kapitel 7.4.3 beschrieben, sind die erfolgreichsten Verfahren zur Bestimmung des Personenparameters und dessen Standardfehlers das Maximum-Likelihood (ML)- und das Expected A Posteriori (EAP)-Verfahren. In dieser Simulationsstudie wurde untersucht, welches dieser Verfahren bei den vorliegenden Items besser geeignet ist, um mit möglichst wenigen Items möglichst gute Schätzungen des Personenparameters zu erzielen. Die Ergebnisse der Testung und die Anzahl der benötigten Test-Items sind auch maßgeblich von den in Kapitel 7.4.5 vorgestellten Verfahren zur Auswahl des nächsten Test-Items abhängig. Deshalb wurden in dieser Simulationsstudie alle möglichen Kombinationen der beiden Schätzverfahren mit den gängigen Verfahren zur Auswahl des nächsten Test-Items verglichen.

Bei der Wahl des ersten Test-Items wurde zu Beginn der Testung davon ausgegangen, dass der Personenparameter einer Versuchsperson 0 ist. Dies führt dazu, dass alle Versuchspersonen mit demselben Item beginnen. Diese Entscheidung erscheint vor dem Hintergrund der in Kapitel 7.3.4.

beschriebenen Reihenfolgeeffekte sinnvoll. Es wurde zudem davon abgesehen, die Wahl des ersten Items zu variieren, um den Zeit- und Rechenaufwand in der Simulationsstudie einzugrenzen.

Bei der Definition des Abbruchkriteriums wurde der Fokus auf den Standardfehler der Schätzung des Personenparameters gelegt. Wie in Kapitel 7.4.4 beschrieben, steht dieser Standardfehler in direktem Zusammenhang mit der Reliabilität der Messung. Um für jede Person eine möglichst zuverlässige Schätzung des Personenparameters zu gewährleisten, wurde die adaptive Testung erst beendet, wenn dieser kleiner als 0,3 war. Die Anzahl der vorgegebenen Items wurde in der Simulationsstudie nach unten hin auf den Wert 2 gesetzt. Dies führt dazu, dass die Testung unabhängig vom Standardfehler mindestens zwei Items umfasst. Die maximale Anzahl von Items, die vorgegeben werden sollen, wurde auf die Anzahl der Items im Item-Pool gesetzt. Dies ermöglicht, dass der Selektionsalgorithmus in Verbindung mit dem Abbruchkriterium in dieser Konfiguration im Extremfall alle verfügbaren Items darbieten kann.

7.5.2 Methode und Ablauf der Simulationsstudie

Die Grundidee der durchgeführten Simulationsstudie ist, die vorliegenden Antwortvektoren der Versuchspersonen aus dem ersten Teil der Studie zu nutzen, um im Rahmen einer Simulation zu prüfen, wie ein adaptiver Test mit einer bestimmten Konfiguration für eine Versuchsperson abgelaufen wäre. Eine ähnliche Strategie verfolgten auch Smits, Cuijpers und van Straten (2011), die untersuchten, welche Konfigurationen für eine adaptive Version der Center for Epidemiologic Studies Depression Scale gewählt werden sollte.

Grundlage der Simulationsstudie sind die Items der Skala Service-Qualität, deren Item-Parameter, wie in Kapitel 7.4.1 dargestellt, als RSM-Modell mit dem R-Paket „eRm“, Version 1.0, bestimmt wurden (Mair & Hatzinger, 2007). Die technische Grundlage der Simulation basiert auf dem R-Skript des Tools Firestar, das von Choi (2009) zur Verfügung gestellt wurde. Sowohl die Schätzverfahren für den Personenparameter als auch die Algorithmen zur Wahl des nächsten Items aus diesem Skript wurden genutzt.

Je nach Konfiguration des adaptiven Tests wurde im Rahmen der Simulation das Item bestimmt, mit dem der Test beginnt. Da die Antwort auf dieses Item im Antwortvektor der Versuchspersonen bereits vorliegt, wurde diese genutzt, um den Personenparameter und dessen Standardfehler zu schätzen. Sobald eine Approximation für den Personenparameter vorlag, wurde der voreingestellte Algorithmus zur Wahl des nächsten Items genutzt, um das Item zu ermitteln, das im Rahmen einer adaptiven Testung als nächstes dargeboten werden würde. Auch für dieses Item wurde die Antwort der Versuchsperson, die bereits im Datensatz vorliegt, ermittelt und basierend darauf erneut der Personenparameter und dessen Standardfehler geschätzt. Dieser Ablauf wurde so lange wiederholt, bis das definierte Abbruchkriterium erfüllt war.

Das Simulationsprogramm protokollierte die finale Theta-Schätzung und deren Standardfehler sowie den Messverlauf, so dass nach dem Durchlauf der Simulation für alle Versuchspersonen dargestellt werden konnte, wie häufig einzelne Items eingesetzt wurden. Darüber hinaus wurde für jede Person festgehalten, wie viele Items zur abschließenden Bestimmung des Personenparameters zum Einsatz kamen.

7.5.3 Ergebnisse der Simulationsstudie

Die Ergebnisse der Simulationsstudie werden in Tabelle 8 dargestellt. Die ersten beiden Spalten zeigen das gewählte Verfahren zur Schätzung des Personenparameters und die damit kombinierten Algorithmen zur Wahl des nächsten Test-Items. In der dritten Spalte ($M_{(N.items)}$) wird angegeben, wie viele Items durchschnittlich pro Versuchsperson beziehungsweise adaptive Testung genutzt wurden. Die vierte Spalte ($M_{(SE)}$) gibt den mittleren Standardschätzfehler der Θ-Schätzung an, der, wie in Formel 9 beschrieben, im Zusammenhang mit der Reliabilität der Θ-Schätzung steht, die in Spalte fünf berichtet wird. Die Spalten sechs und sieben enthalten die Korrelation des Θ-Wertes des adaptiven Tests mit dem ermittelten Θ-Wert basierend auf dem vollständigen Antwortvektor ($r_{(\Theta.cat,\ \Theta.all)}$) und dem Skalenmittelwert ($r_{(\Theta.cat,\ M(Skala))}$), der im Rahmen der klassischen Testauswertung als Gesamtschätzer genutzt wird.

Tabelle 8 Überblick über die Ergebnisse der Simulationsstudie

Bestimmung von Theta	Wahl des nächsten Items	$M_{(N.Items)}$	$M_{(SE)}$	Reliabilität	$r_{(\Theta.cat,\ \Theta.all)}$	$r_{(\Theta cat,\ M(Skala))}$
EAP	MFI	19,77	0,293	.914	.811	.837
EAP	MLWI	16,92	0,293	.914	.806	.831
EAP	MPWI	16,89	0,293	.914	.806	.831
EAP	MEI	17,45	0,293	.914	.807	.831
EAP	MEPV	19,54	0,293	.914	.810	.824
EAP	MEPWI	16,89	0,293	.914	.806	.831
ML	MFI	19,39	0,293	.914	.807	.833
ML	MLWI	16,53	0,293	.914	.800	.825
ML	MPWI	16,49	0,293	.914	.801	.825
ML	MEI	17,06	0,293	.914	.801	.824
ML	MEPV	19,21	0,292	.914	.804	.818
ML	MEPWI	16,49	0,293	.914	.801	.825

Die durchschnittlichen Standardschätzfehler der Θ-Schätzung fallen für alle simulierten Kombinationen ähnlich hoch aus. Dies bestätigt, dass das auf dem Standardfehler der Θ-Schätzung basierende Abbruchkriterium für die adaptive Testung in nahezu allen Konfigurationen erreicht bzw. unterschritten wurde. An den durchweg sehr hohen Korrelationen zwischen dem Θ-Wert aus der adaptiven Testung und den Validitätskriterien kann festgemacht werden, dass es eine hohe Übereinstimmung zwischen der adaptiven und nicht adaptiven Messung von Service-Qualität gibt. Erwartungskonform unterscheiden sich die verschiedenen Konfigurationen darin, wie viele Items durchschnittlich benötigt wurden, um eine ausreichend genaue Θ-Schätzung zu erzielen. Die Spanne reicht hier von 16,49 Items (ML + MPWI / ML + MEPWI) bis 19,77 (EAP + MFI). Neben der Betrachtung dieses Mittelwerts lohnt es sich, die Verteilung der Anzahl der benötigten Items zu visualisieren. Eine entsprechende Darstellung wurde für alle verglichenen Konfigurationen der Simulationsstudie erstellt. Abbildung 28 zeigt, exemplarisch für die Kombination EAP und MPWI, die Verteilung der Anzahl der eingesetzten Items als Boxplot.

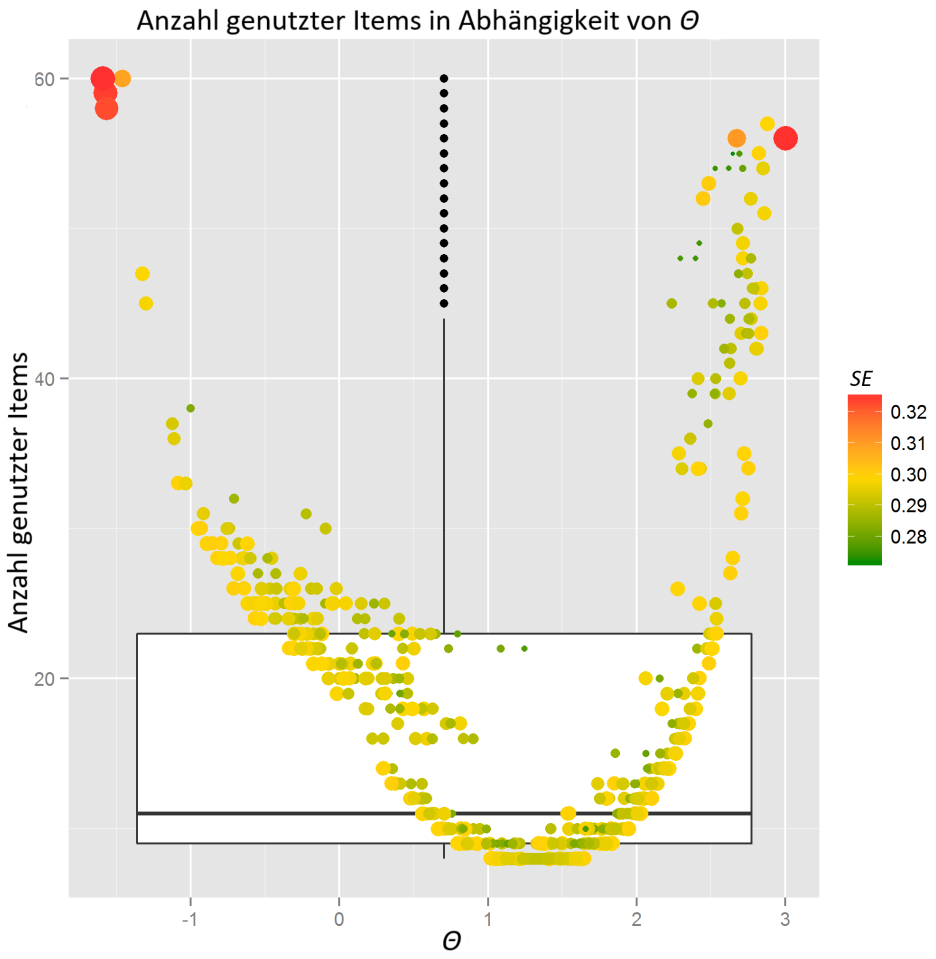


Abbildung 28 Darstellung der Anzahl genutzter Items in Abhängigkeit der Ausprägung des Personenparameters Θ und farb- und größenkodierte Darstellung der jeweiligen Standardschätzfehler für die Simulationsbedingung EAP und MPWI

Der im Boxplot dargestellte Median zeigt, dass bei 50 % der Fälle weniger als 12 Items für eine ausreichend präzise Bestimmung des Personenparameters erforderlich waren. Die Anordnung der farbigen Kreise in Abbildung 28 entlang der y-Achse stellen die Anzahl der benötigten Items in Abhängigkeit des Personenparameters Θ (x-Achse) dar. Die Größe und Farbe der Kreise kodiert den mit der Θ-Schätzung verbundenen Standardschätzfehler. An der nahezu parabelförmigen Anordnung dieser Kreise wird deutlich, dass in den Extrembereichen des Personenparameters besonders viele Items eingesetzt wurden. Insbesondere für $\Theta < -1{,}5$ konnte trotz der Nutzung aller verfügbaren Items der als Abbruchkriterium definierte Standardschätzfehler nicht unterschritten werden.

Um die Zusammenhänge zwischen der Θ-Schätzung des adaptiven Tests und dem Skalenmittelwert genauer untersuchen zu können, wurden für alle untersuchten Konfigurationen Streudiagramme erstellt. Die mit der Kombination EAP und MPWI verknüpfte Korrelation von $r = .83$ zwischen dem Personenparameter des simulierten adaptiven Tests und dem Skalenmittelwert, der als Validitätskriterium genutzt wurde, ist in Abbildung 29 exemplarisch dargestellt.

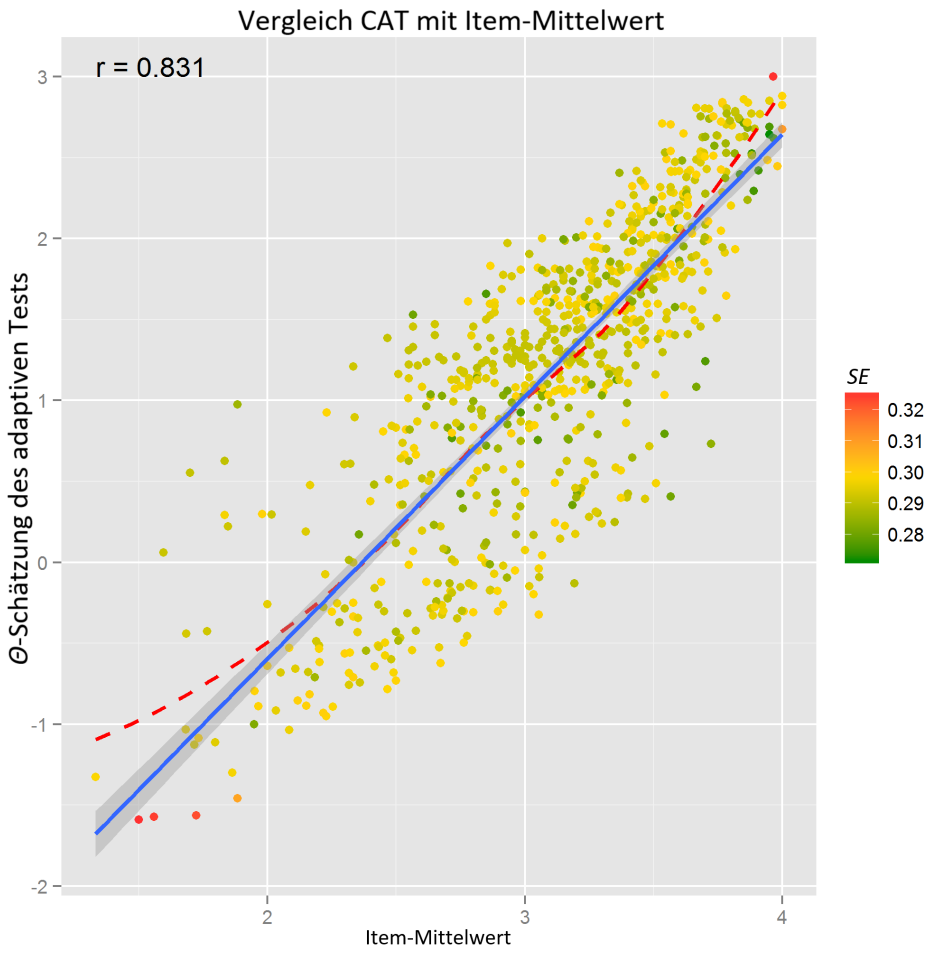


Abbildung 29 Zusammenhang zwischen dem Θ-Schätzer des adaptiven Tests und dem Skalenmittelwert in der Konfiguration EAP und MPWI

Die Farbe der Punkte in Abbildung 29 kodiert den Standardfehler der Θ-Schätzung. Zudem enthält die Abbildung in blauer Farbe eingezeichnet eine Regressionsgerade mit einem 95 prozentigen Konfidenzband darum, das als graue Fläche dargestellt ist. Das Ergebnis einer lokalen loess-Regression wurde als rote gestrichelte Linie eingezeichnet (Wickham, 2016). Der nahezu perfekte lineare Zusammenhang zeigt, dass die Ergebnisse der adaptiven Testung mit denen der nicht adaptiven Testung sehr gut übereinstimmen.

Wie in Kapitel 7.4.5 beschrieben, kann die Auswahl des nächste Items nach Algorithmen, die sich streng an psychometrischen Kriterien orientieren, dazu führen, dass bestimmte Items sehr häufig, andere extrem selten dargeboten werden. Um in dieser Simulationsstudie zu prüfen, ob bestimmte Kombinationen aus Θ-Schätzverfahren und Selektionsalgorithmus zu auffälligen Mustern in der Nutzung der Items führten, wurde die Nutzung der Items für jede simulierte Konfiguration untersucht. Abbildung 30 zeigt die Nutzung der Items exemplarisch für die Konfiguration EAP und MPWI. Auf der x-Achse sind die 60 Items dargestellt. Der auf der y-Achse dargestellte Prozentwert errechnet sich als Anteil der Häufigkeit, mit der ein Item dargeboten wurde, an der Gesamtanzahl von Items, die über alle Versuchspersonen eingesetzt wurden.

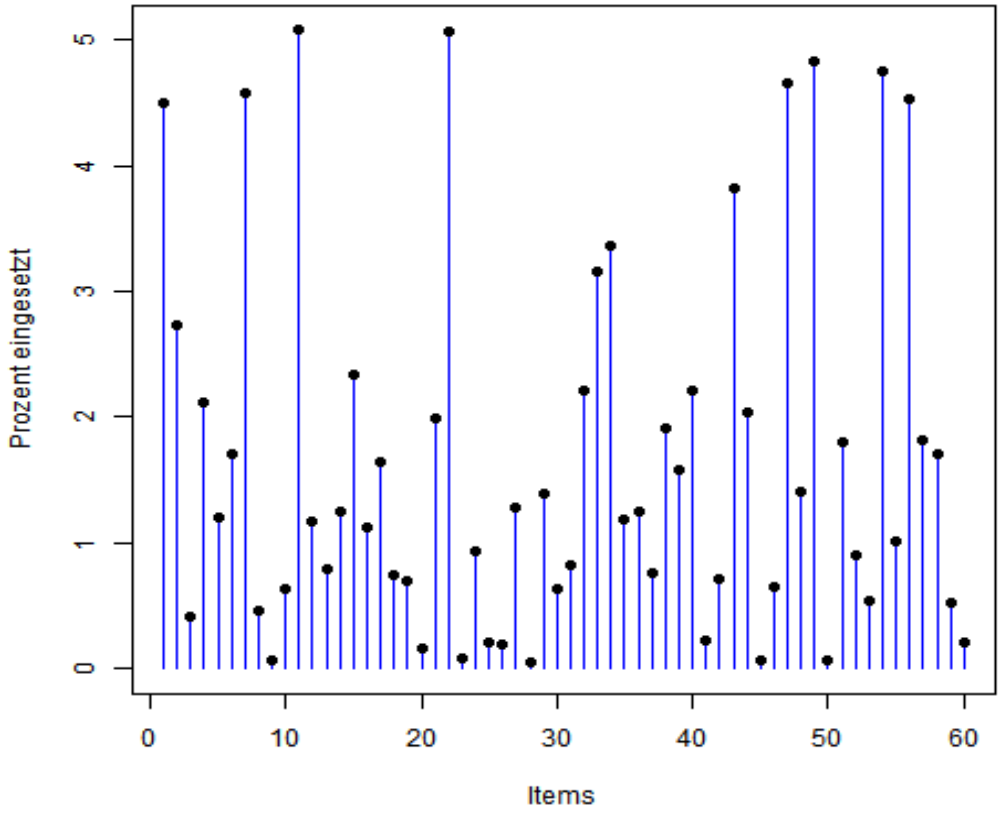


Abbildung 30 Relative Häufigkeit der Darbietung der einzelnen Items für die Konfiguration EAP und MPWI

Wie zu erwarten war, visualisiert Abbildung 30, dass in der Konfiguration EAP in Verbindung mit dem Selektionsalgorithmus MPWI Beispielsweise Item 11 und 21 sehr häufig und die Items 9, 28, 45 und 50 sehr selten auswählt wurden.

7.5.4 Diskussion und Schlussfolgerungen aus der Simulationsstudie

Das zentrale Anliegen dieser Simulationsstudie war herauszufinden, welche Kombination aus Θ-Schätzverfahren und Selektionsalgorithmus für das nächste Item, bei dem gegebenen Item-Pool, bestehen aus den Items der Skala Service-Qualität, durchschnittlich mit möglichst wenig Items pro Testung auskommt. Dabei galt als Rahmenbedingung, dass die Schätzung des Personenparameters möglichst präzise erfolgt, was durch die Definition des Abbruchkriteriums anhand des Standardfehlers der Θ-Schätzung sichergestellt wurde. Zudem wurde korrelativ untersucht, wie gut die Ergebnisse der adaptiven Testung mit den Ergebnissen der nicht adaptiven Erfassung von Service-Qualität übereinstimmen. Die Ergebnisse aus Tabelle 8 zeigen, dass die adaptive Testung in allen verglichenen Konfigurationen zu einer reliablen Schätzung des Personenparameters führte. Setzt man die Genauigkeit der Θ-Schätzung wie in dieser Simulationsstudie auf einen Standardfehler von < 0.3 fest, so ergeben sich über alle Kombinationen aus Θ-Schätzverfahren und Auswahlalgorithmus für das nächste Item hohe Korrelationen zwischen dem Ergebnis der adaptiven Testung und dem Skalenmittelwert der Skala zur Erfassung von Service-Qualität. Abbildung 29, die diesen Zusammenhang exemplarisch visualisiert, zeigt, dass die Annahme eines linearen Zusammenhangs berechtigt ist und die Residuen über den gesamten Vorhersagebereich ähnlich groß ausfallen. Die hohen korrelativen Übereinstimmungen zwischen den Testergebnissen der adaptiven und nicht adaptiven Testung können als Indikator für die Validität der adaptiven Erfassung von Service-Qualität gewertet werden.

Der Vergleich der beiden Schätzverfahren für den Personenparameter in Tabelle 8 zeigt, dass die Ergebnisse relativ ähnlich ausfallen. Betrachtet man die durchschnittlich benötigte Anzahl von Items, so kommt das ML-Verfahren zur Schätzung des Personenparameters mit etwas weniger Items aus. Fokussiert man auf die Korrelation mit dem genutzten Validitätskriterium, zeigen sich leichte Vorteile für das EAP-Verfahren. Da dieses, wie in Kapitel 7.4.3 beschrieben, weniger rechenintensiv ist und auch bei konstanten Antwortvektoren eine Schätzung liefert, wird das EAP-Verfahren für die Entwicklung dieses adaptiven Tests favorisiert.

Ein Vergleich der verschiedenen Algorithmen zur Selektion des nächsten Test-Items zeigt, dass diese zu relativ ähnlichen Ergebnissen führen. Es ergeben sich lediglich geringe Unterschiede in der durchschnittlich benötigten Anzahl von Items. Vor dem Hintergrund möglichst hoher Ökonomie schneiden die Verfahren MPWI und MEPWI am besten ab. Wie Choi et al. (2012) mathematisch nachweisen konnten, unterscheiden sich die Ergebnisse des MPWI- und

des MEPWI-Algorithmus in Tabelle 8, unabhängig des genutzten Schätzverfahrens für den Personenparameter, nicht. Da das MPWI-Verfahren viel einfacher und weniger rechenintensiv ist, kann dieses als Ergebnis der Simulationsstudie für die praktische Entwicklung eines adaptiven Tests für Service-Qualität empfohlen werden.

Betrachtet man die verschiedenen Konfigurationen, die in dieser Simulationsstudie eingesetzt wurden, vor den Hintergrund der Optimierung der Testlänge, so kann an der Anzahl der durchschnittlich eingesetzten Items erkannt werden, dass verglichen mit der nicht adaptiven Erfassung von Service-Qualität mit 61 Items deutlich weniger Items zur Schätzung des Personenparameters ausreichend sind. Die Vorgabe aller Fragebogenitems hat somit in vielen Fällen nur geringen zusätzlichen Informationswert. Betrachtet man Abbildung 28 exemplarisch, wird deutlich, dass im mittleren Bereich von Θ, in dem die meisten Beobachtungen zu verzeichnen sind, meist weniger als zehn Items zur präzisen Schätzung des Personenparameters ausreichend waren. Vergleicht man Abbildung 28 aus dieser Simulationsstudie mit Abbildung 27 von Embretson und Reise (2000), so ergibt sich ein Widerspruch. Embretson und Reise (2000) zeigen in ihrer Abbildung, dass die Messgenauigkeit adaptiver Tests in den Extrembereichen genauso groß ist wie im mittleren Bereich. In der vorliegenden Simulationsstudie wurde beobachtet, dass die Präzision der Messung in den Extrembereichen der latenten Dimension nachlässt. In der Simulation von Embretson und Reise (2000) war das adaptive Verfahren auf eine bestimmte Anzahl von Items fixiert. In der vorliegenden Simulationsstudie wurde das Abbruchkriterium, um eine möglichst hohe Messgenauigkeit zu erzielen, so gewählt, dass maximal alle verfügbaren Items eingesetzt werden konnten. Der zentrale Unterscheid zwischen den beiden Studien besteht im genutzten Item-Pool. Während Embretson und Reise (2000) mit fiktiven Items, deren Schwierigkeit den gesamten Bereich der latenten Dimension ideal abdeckten, arbeiteten, nutzte die vorliegende Simulationsstudie die die Items der Skala Service-Qualität, die, wie in Anhang F zu sehen ist, den mittleren Bereich der latenten Dimension besser abdecken als die Randbereiche. Die Annahme, dass adaptive Tests in den Extrembereichen der Merkmalsausprägung grundsätzlich zu präziseren Messergebnissen führen, muss deshalb relativiert werden. Die hier durchgeführte Simulationsstudie zeigt, dass diese Messgenauigkeit von adaptiven Tests in den Extrembereichen der latenten Dimension stark vom verfügbaren Item-Pool abhängig ist.

Wie weitere, hier nicht im Detail dargestellte Simulationen zeigten, reduziert sich die Anzahl durchschnittliche benötigter Items deutlich, wenn einzelne Fälle mit extremen Antwortvektoren aus der Simulation ausgeschlossen werden. In künftigen Studien sollten deshalb extreme Messwerte dahingehend untersucht werden, ob sich dahinter tatsächlich extreme Merkmalsausprägungen verbergen. Sollte dies nicht der Fall sein, wird empfohlen, diese Fälle von der Simulation auszuschließen.

Eine Grundannahme, die hinter dieser Art der Simulation steht, ist, dass die Antworten auf die Items, die in der nicht adaptiven, linearen Testung gegeben wurde, identisch ausgefallen

wären, wenn die Items in einer anderen, von der Konfiguration des adaptiven Tests abhängigen Reihenfolge dargeboten werden würden. Diese Annahme ist, wie in Kapitel 7.3 dargestellt wurde, anzweifelbar. Solange jedoch keine klaren Hypothesen und Befunde zu möglichen systematischen Effekten der Item-Abfolge vorliegen, erscheint der Ansatz dieser Simulationsstudie sinnvoll, da ein bedeutender Erkenntniszugewinn möglich ist. Für die künftige Forschung auf diesem Gebiet wäre es interessant zu untersuchen, ob und wie sich das Antwortverhalten von Versuchspersonen bei adaptiver und nicht adaptiver Darbietung der Items unterscheidet. Sollte die Art der Darbietung keinen systematischen Effekt auf das Antwortverhalten haben, würde dies für die in dieser Simulation genutzte Strategie sprechen.

In dieser Simulationsstudie wurde als Kriterium geprüft, ob die Ergebnisse einer klassischen Testung durch einen adaptiven Test repliziert werden können. Dies ist ein mögliches Validitätskriterium, das sich anbietet, weil es leicht verfügbar ist. Diese Art der Validierung sollte in künftigen Studien um weitere Kriterien ergänzt werden. Es sollte beispielsweise geprüft werden, ob die Ergebnisse der adaptiven Testung auch mit Fremdeinschätzungen der Service-Qualität, zum Beispiel durch andere Akteure im Markt oder Kundinnen und Kunden, zusammenhängen. Neben diesen Kriterien sollte auch die prädiktive Validität zu Verhaltensmaßen und der Zusammenhang mit Kundenzufriedenheit berücksichtigt werden. Die konvergente und diskriminante Validität zu anderen Konstrukten und Messmethoden könnte im Rahmen des Multitrait-Multimethod-Ansatzes untersucht werden (Campbell & Fiske, 1959; Eid & Diener, 2006).

Eine Herausforderung von adaptiven Testverfahren besteht darin, dass nur eine Auswahl aus allen Items dargeboten wird. Diese selektive Darbietung kann dazu führen, dass Items, die wichtige inhaltliche Aspekte des zu messenden Konstrukts erfassen, nicht dargeboten werden. In solchen Fällen kann angezweifelt werden, dass der adaptive Test immer noch ein valides Maß für das zu erfassende Konstrukt darstellt. Auch in dieser Simulationsstudie wurden die Items der Skala Service-Qualität, wie in Abbildung 30 exemplarisch dargestellt, unterschiedlich oft genutzt. Ob dies tatsächlich ein Problem darstellt und die Qualität der Messergebnisse mindert, sollten weitere Validierungsstudien untersuchen. Um sicherzustellen, dass aus allen inhaltlich wichtigen Bereichen des Konstrukts eine Mindestanzahl von Items dargeboten wird, könnten Techniken wie „Content Balancing“ einsetzt werden (Cheng et al., 2007; Han, 2018).

Die Übertragbarkeit der Ergebnisse dieser Simulationsstudie auf andere Testverfahren mit anderen Test-Items und anderen zugrunde liegenden empirischen Daten ist nur eingeschränkt möglich. Pauschale Empfehlungen für zum Beispiel einen Algorithmus zur Bestimmung des nächsten Items oder ein Abbruchkriterium können nicht gegeben werden. Die Erfahrungen aus verschiedenen Simulationsdurchläufen sprechen dafür, dass solche Simulationen für jedes diagnostische Instrument gesondert durchgeführt werden sollten.

Die Nutzung von Simulationsstudien zur Prüfung möglicher Konfigurationen von adaptiven Tests stellt einen ökonomischen Weg dar, um durch entsprechende Visualisierungen und Statistiken genauere Einblicke in den Ablauf adaptiver Test zu gewinnen. Sie ermöglichen darüber hinaus eine genaue Analyse, wie sich verschiedene Einstellungsparameter adaptiver Tests auf den Verlauf und die Genauigkeit der Erfassung des latenten Konstrukts auswirken.

Wenn wie in dieser Studie nach einer Konfiguration gesucht wird, die das Ergebnis einer klassischen linearen Testung unter Berücksichtigung aller Items möglichst gut repliziert und dabei mit möglichst wenigen Items auskommt, sprechen die Ergebnisse aus Tabelle 10 für die Kombination aus dem Schätzverfahren EAP und dem Auswahlalgorithmus MPWI. Auch der genauere Blick auf diese Kombination, der in Abbildung 28 und Abbildung 29 geboten wird, spricht für diese Kombination. Für die praktische Umsetzung des webbasierten adaptiven Tests für Service-Qualität wird diese Konfiguration deshalb empfohlen.

7.6 Technische Grundlage für webbasierte adaptive Tests

Die meisten bislang entwickelten adaptiven Tests wurden als eigenständige Software entwickelt, die in der Regel offline auf einem Computersystem mit einem bestimmten Betriebssystem ausgeführt wird. Chien, Wang, Huang, Lai und Chow (2011) entwickelten ein Visual Basic Application (VBA)-Modul, das in Verbindung mit Microsoft Excel funktioniert. Diese Lösung konnten sie auch über das Internet zur Verfügung stellen. In dieser Arbeit wurde das Ziel verfolgt, eine Lösung zu entwickeln, die auf zeitgemäßen Webtechnologien basiert, so dass die Erhebung des fertigen adaptiven Tests webbasiert und damit mit jedem internetfähigen Endgerät, das über einen modernen Browser verfügt, durchgeführt werden kann. Diese Herangehensweise erleichtert die Anwendung adaptiver Tests, denn es werden keine spezifischen Anforderungen an das Computersystem der Testpersonen gestellt. Sobald eine Verbindung zum Internet und ein aktueller Browser zur Verfügung stehen, kann die Testung ortsunabhängig durchgeführt werden.

Als Grundlage für die webbasierte Umsetzung des entwickelten adaptiven Tests wurde ein Serversystem mit dem Betriebssystem SUSE Linux Enterprise Server (Version 11) gewählt. Als Webserver kommt der freie und quelloffene Apache HTTP-Server (Version 2.2.12) zum Einsatz, der als meistgenutzter Webserver im Internet gilt. Um die benötigten dynamischen Webseiten zu generieren, wurde eine Kombination aus HTML und der serverseitigen Skriptsprache PHP (Version 5.4.20) genutzt. Für die Berechnungen, die z. B. zur Bestimmung des Personenparameters und der Festlegung des nächsten Items notwendig sind, wurde das frei verfügbare Statistikprogramm R (Version 3.0.2) eingesetzt.

7.7 Ablauf des entwickelten webbasierten adaptiven Tests

Der entwickelte webbasierte adaptive Test zur Erfassung von Service-Qualität ist so aufgebaut, dass den Versuchspersonen die Items einzeln und nacheinander auf dem Bildschirm dargeboten werden. Die dargebotenen Items müssen nacheinander beantwortet werden; es ist nicht möglich, Items zu überspringen. Nachdem ein Item beantwortet wurde, gibt es keine Möglichkeit, zu zurückliegenden Items zurück zu navigieren und bereits gegebene Antworten zu verändern. Diese klassische Art der Darbietung adaptiver Tests wurde vor dem Hintergrund der in Kapitel 7.3.5 beschrieben Herausforderungen favorisiert.

Nach Aufruf der Website erfolgt eine Begrüßung, gefolgt von Informationen und Instruktionen zum bevorstehenden Test. Anschließend wird die Testperson über die Eingabe eines gültigen Zugangscodes authentifiziert und am System angemeldet. Nach einem erfolgreichen Login wird durch PHP das Programm R mit dem Auftrag, das Skript „adaptfirst.R“ auszuführen, gestartet. Dabei übergibt PHP die Benutzerkennung an R. R legt nach dem Einlesen der Konfigurationsdatei „config.R“ für die aktuelle Testung einen sogenannten Workspace an und speichert diesen, mit der Benutzerkennung als Element des Dateinamens ab. In diesem Workspace werden die benötigten Funktionen und alle Informationen für die Testung erfasst und gespeichert. R legt bei diesem ersten Aufruf gemäß der vorgegebenen Konfiguration fest, mit welchem Item der Test beginnt. Nachdem dieses R-Skript ausgeführt wurde, gibt R an PHP zurück, mit welchem Item der Test beginnen soll. PHP generiert anhand dieser Rückmeldung die erste Seite des adaptiven Tests und stellt darauf das erste Item dar. Über das Anklicken von Radiobuttons einer vierstufigen Antwortskala, deren Extreme verbal mit „trifft nicht zu“ und „trifft voll zu“ verankert wurden, wird angegeben, wie sehr die Aussage des dargebotenen Items zutrifft. Durch das Anklicken der Schaltfläche mit der Beschriftung „weiter“ wird die Auswahl bestätigt und das HTML-Formular abgeschickt. Im nächsten Schritt prüft PHP die eingegangene Antwort, startet R und übergibt die Antwort in Verbindung mit der Benutzerkennung. R verarbeitet die Antwort, schätzt den aktuellen Personenparameter und Standardfehler (siehe Kapitel 7.4.3) und prüft, ob der Test beendet werden soll (siehe Kapitel 7.4.4). Sind die Abbruchkriterien nicht erreicht, bestimmt R das nächste Item (siehe Kapitel 7.4.5), das vorgegeben werden soll, und gibt dieses an PHP zurück. In diesem Fall generiert PHP eine weitere Fragebogenseite, auf der die Antwort auf dieses Item erfasst wird. Kommt R zu dem Ergebnis, dass eines der Abbruchkriterien erfüllt ist, wird der endgültige Personenparameter und der Standardfehler bestimmt und an PHP zurückgemeldet. Auf der abschließend dargebotenen Seite „end.php“ wird das Testergebnis in Form eines Personenparameters und eines entsprechenden Standardfehlers zurückgemeldet. Zusätzlich wird durch R eine Grafik erzeugt, die den Messverlauf darstellt und in diese Seite eingebunden werden kann. Abbildung 31 zeigt exemplarisch eine solche Grafik.

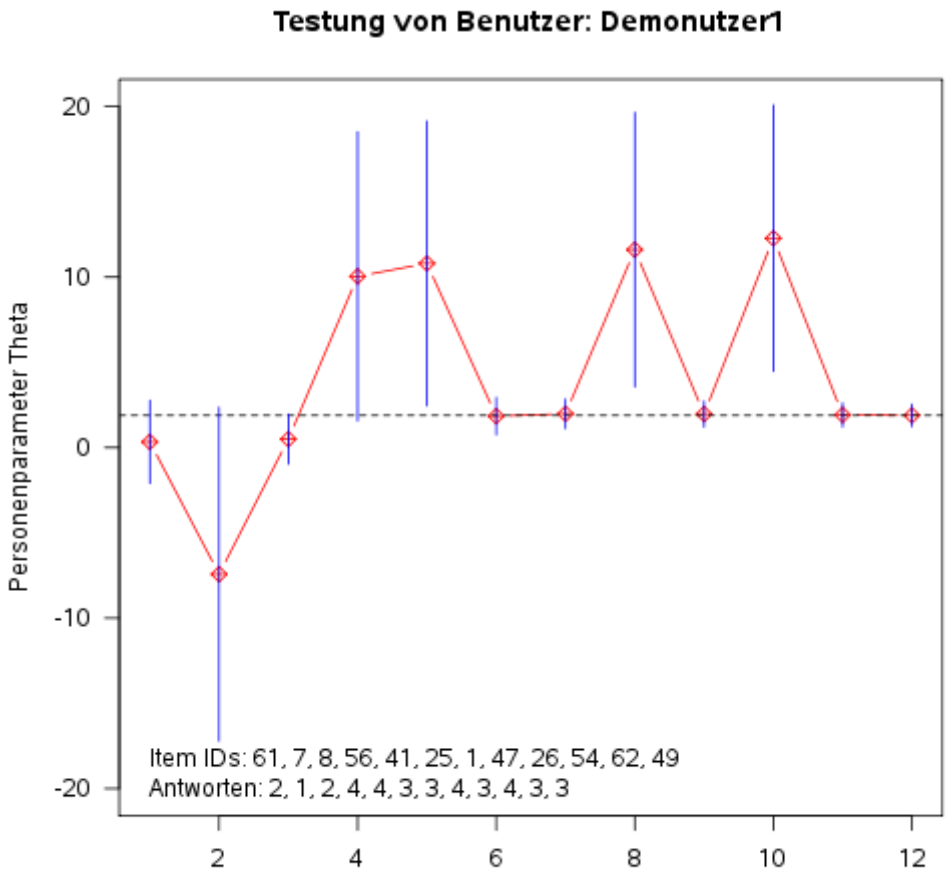


Abbildung 31 Exemplarische Grafik zur Darstellung des Testverlaufs

Die x-Achse dieser Grafik symbolisiert den Verlauf der Messung. In der dargestellten Beispielgrafik wurden 12 Items dargeboten. Auf der y-Achse wird die Ausprägung von Θ dargestellt. Die roten Kreise in der Grafik stellen den Punktschätzer für Θ nach der Beantwortung des jeweiligen Items dar. Die senkrechten blauen Linien zeigen die Breite des auf dem Standardfehler der Schätzung beruhenden Konfidenzintervalls an, in dem mit 95-prozentiger Wahrscheinlichkeit der tatsächliche Θ-Wert liegt. Zusätzlich zu diesen grafischen Elementen werden oberhalb der x-Achse die Item-Nummern der dargebotenen Items dargestellt und die von Nutzer ausgewählten Antworten zu diesen Items ausgegeben.

8 Abschlussdiskussion

In der folgenden Abschlussdiskussion werden zunächst die zentralen Ergebnisse dieser Arbeit dargestellt und diskutiert. Anschließend wird erörtert, welche Tragweite die Ergebnisse haben und in welchem Maß sie generalisiert werden können. Die Anwendungsmöglichkeiten der Skala zur Erfassung von Service-Qualität und des adaptiven Tests zur Erfassung von Service-Qualität sowie deren Vor- und Nachteile werden zusammengefasst. Weiterhin werden die Erfahrungen, die bei der Praxiserprobung des adaptiven Tests zur Erfassung von Service-Qualität gesammelt wurden, dargestellt und aufgezeigt, welche kritischen Aspekte in der Praxis berücksichtigt werden sollten. Abschließend wird zusammengefasst, welche Implikationen sich für die künftige Forschung ableiten lassen, und im Rahmen eines Ausblicks skizziert, welche Relevanz die ökonomische Erfassung von Service-Qualität durch adaptive Testverfahren künftig haben könnte.

8.1 Zusammenfassung der zentralen Ergebnisse

In Teil 1 dieser Arbeit wurde ein diagnostisches Instrument zur Erfassung von Service-Qualität entwickelt. Durch einen Vortest und den Einsatz dieses Instruments wurden Daten gewonnen, die genutzt wurden, um zu prüfen, ob die psychometrischen Eigenschaften der entwickelten Items zufriedenstellend sind. Die Ergebnisse dieser Analysen ergaben minimalen Anpassungsbedarf und führten zu einem Befragungsinstrument, dessen intendierte faktorielle Struktur mit den angestrebten vier Subdimensionen mittels einer konfirmatorischen Faktorenanalyse untersucht und bestätigt wurden. Basierend auf dem aktuellen Forschungsstand wurde ein Rahmenmodell, das die Zusammenhänge zwischen Service-Qualität, Kundenzufriedenheit und dem Erfolg der Organisation als Mediationsmodell beschreibt, hergeleitet. Die Hypothesen dieses Modells wurden in ein Strukturgleichungsmodell überführt, das einen zufriedenstellenden Modellfit aufwies. Wie angenommen, konnte gezeigt werden, dass Service-Qualität über die vier gewählten Subfacetten gemessen werden kann und im Zusammenhang mit Kundenzufriedenheit steht. Auch der angenommene Zusammenhang zwischen Kundenzufriedenheit und dem Erfolg der Organisation ließ sich empirisch bestätigen. Die gefundenen Effekte decken sich, auch hinsichtlich ihrer Größenordnung, gut mit den Ergebnissen vergleichbarer Studien, was darauf hindeutet, dass das entwickelte Befragungsinstrument zur Messung von Service-Qualität aus der Binnenperspektive einer Organisation geeignet ist. Zudem kann die empirische Bestätigung des Rahmenmodells als ein weiterer Hinweis für die Gültigkeit der Annahmen der Service-Profit Chain gewertet werden.

Im zweiten Teil dieser Arbeit wurde untersucht, wie ein adaptiver Test zur Erfassung von Service-Qualität entwickelt werden kann. Dabei wurde zunächst dargestellt, welche Arten von adaptiven Tests unterschieden werden können, und festgelegt, dass eine internetbasierte Erfassung als zeitgemäße Lösung favorisiert wird. Weil es bei der Entwicklung und Umsetzung adaptiver Tests einige Entscheidungen zu treffen gilt, wie zum Beispiel der Personenparameter oder das nächste Item bestimmt werden sollen, wurde eine Simulationsstudie genutzt, um unterschiedliche Konfigurationen zu vergleichen. Diese Herangehensweise erwies sich als nützlich und konnte zeigen, wie sich unterschiedliche Konfigurationen auf die Testlänge und das Testergebnis auswirken. Die Ergebnisse der Simulationsstudie lieferten detailliert Aufschluss darüber, wie ein adaptives Testsystem konfiguriert sein sollte, um mit möglichst wenigen Fragen ein möglichst exaktes Messergebnis zu liefern. Diese Erkenntnisse wurden bei der Entwicklung und Implementierung des internetgestützten adaptiven Tests zur Erfassung von Service-Qualität genutzt.

8.2 Theoretische und methodische Einschränkungen

Der erste Teil dieser Arbeit, in dem die Bedingungen und Konsequenzen von Service-Qualität untersucht wurden, schließt mit einer Diskussion ab, in bereits auf methodische Einschränkungen hingewiesen wurde (s. Kap. 6.2), die bei der Interpretation und Generalisierung der Ergebnisse zum Rahmenmodell dieser Studie berücksichtigt werden sollten. Dabei wurde die Operationalisierung der Kundenzufriedenheit und des Erfolgs von Organisationen, die sehr ökonomisch erfolgte, als kritisch betrachtet. Zudem sollte, da keine echte Zufallsstichprobe zugrunde liegt, geprüft werden, ob sich die gefundenen Ergebnisse auch bei anderen, möglichst repräsentativen Stichproben replizieren lassen. Fehlende Werte in den Daten wurden vermehrt bei Items beobachtet, die in der linearen Darbietungsform der Skala zur Erfassung von Service-Qualität weiter hinten erfasst wurden. Dies kann als ein Indikator dafür gewertet werden, dass einige Versuchspersonen die Befragung vorzeitig abgebrochen haben. Sollte es systematische Zusammenhänge zwischen der Ausprägung der Service-Qualität und der Tendenz, die Befragung vorzeitig abzubrechen, geben, so würde dies die Aussagekraft der Daten ebenfalls einschränken.

Das Rahmenmodell dieser Arbeit geht, wie in der Theorie zur Service-Profit Chain angenommen, von kausalen Zusammenhängen aus. Dabei wird Service-Qualität als Ursache oder Ausgangspunkt betrachtet, die sich vermittelt über Kundenzufriedenheit auf den Erfolg von Organisationen auswirkt. Auch wenn die in dieser Studie berichteten Pfadkoeffizienten des Strukturgleichungsmodells signifikante Zusammenhänge zwischen den Konstrukten identifizieren konnte, ist eine kausale Interpretation dieser Zusammenhänge nicht zulässig. Eine experimentelle Herangehensweise, bei der das Niveau von Service-Qualität bewusst variiert

wird, oder die Erfassung der Konstrukte zu mehreren Zeitpunkten könnten Aufschluss über die Frage der Kausalität geben.

Im zweiten Teil dieser Arbeit wurde berichtet, wie die Entwicklung eines adaptiven Tests zur Erfassung von Service-Qualität durchgeführt werden kann. Ein zentraler Mehrwert des adaptiven Tests zur Erfassung von Service-Qualität ist die deutlich verkürzte Testdauer. Während für die Bearbeitung der Skala zur Erfassung von Service-Qualität 61 Fragen beantwortet müssen, kann die adaptive Variante des Verfahrens, je nach Merkmalsausprägung, bereits mit 10 bis 20 Fragen eine nahezu gleich gute Schätzung der Service-Qualität erfolgen. Diese enorme Verkürzung der Testlänge führt zu einer deutlich höheren Ökonomie des Verfahrens (Choi, Reise et al., 2010). Die Ökonomie des adaptiven Tests ist darüber hinaus hoch, weil die Durchführung und Auswertung vollständig automatisiert wurde und damit, verglichen mit klassischen Testverfahren, die Kosten für die Personen, die den Test instruieren, durchführen und auswerten, entfallen. Auch die Speicherung und Verarbeitung der Antwortdaten und das Generieren von individuellen Rückmeldungen und Berichten für verschiedene Bereiche wurde vollständig automatisiert. So können Fehler, die bei einer manuellen Weiterverarbeitung der Daten entstehen können, vermieden und weitere Kosten eingespart werden. Angesichts der Möglichkeit, Service-Qualität zeit- und ortsunabhängig auf nahezu jedem internetfähigen Gerät zu erfassen, ist davon auszugehen, dass die Bearbeitung für die Versuchspersonen komfortabel ist, so dass mit einer höheren Beteiligung und einer guten Datenqualität zu rechnen ist.

Wenn in der adaptiven Testvariante in der Einzelfalldiagnostik nur ein Bruchteil des gesamten Item-Pools genutzt wird, so wird objektiv weniger Information und damit das Konstrukt nicht in seiner vollen Breite erfasst. Dies könnte mit einer geringeren Validität der Messung einhergehen. Insbesondere bei einer Einzelfalldiagnostik, die die Absicht verfolgt, gezielte Interventionen zu planen, kann sich dies nachteilig auswirken. Für die wissenschaftliche Auseinandersetzung mit dem Thema Service-Qualität, in der es üblicherweise um Gruppendiagnostik geht, stellt dies kein Problem dar. Zudem bietet die adaptive Erfassung des Konstrukts die Möglichkeit, bei gleichem zeitlichem Aufwand mehrere Personen zu befragen, um einen breiteren diagnostischen Zugang zur Service-Qualität einer Organisation zu erlangen. Wenn Service-Qualität als ein wichtiger Indikator für das Management von Organisationen laufend erfasst und überwacht werden soll, überwiegen die Vorteile der adaptiven Erfassung.

Da ein adaptiver Test in der Regel aus einem großen Item-Pool die Items auswählt, die für die Diagnostik im aktuellen Fall den höchsten Informationswert haben, entsteht aus Sicht der Versuchspersonen bei jeder Erfassung des Konstrukts ein individueller Test. Die so entstehenden Testabläufe können als Paralleltests betrachtet werden und bei wiederholter Messung des Konstrukts Übungs- und Erinnerungseffekte minimieren.

Wie bereits in Kapitel 7.3.2 beschrieben, kann davon ausgegangen werden, dass diese Eigenschaft adaptiver Tests sich günstig auf die Motivation der Versuchspersonen auswirkt. Bei

häufiger Messung entsteht bei den Versuchspersonen nicht der Eindruck, laufend die gleichen Fragen beantworten zu müssen. Weil der Selektionsalgorithmus Fragen ausgewählt, die einen hohen Informationswert haben, sollten den Versuchspersonen keine Fragen gestellt werden, deren Antwort bereits aus der Beantwortung zurückliegender Items hätte abgeleitet werden können. Die Bearbeitung des adaptiven Tests sollte der Interaktion mit einer Person, die gezielt relevante Fragen stellt, nahekommen und deshalb als angenehmer empfunden werden.

Da die Grundlage für adaptive Tests die Item-Parameter sind, die in der Regel mit Modellen der probabilistischen Testtheorie, bestimmt werden, erfordert die Entwicklung eines solchen Testverfahrens eine vertiefte Auseinandersetzung mit den Ansätzen der Item-Response-Theorie. Insbesondere bei polytomen Antwortformaten, die vom einfachen Rasch-Modell nicht mehr abgebildet werden können, müssen auch komplexere IRT-Modelle bedacht werden. Diese vertiefte Auseinandersetzung mit den psychometrischen Eigenschaften der Items und Antwortformate führt zu einer intensiven Auseinandersetzung mit den einzelnen Items. Ungünstige Items, die zum Beispiel extrem leichte Schwierigkeitsparameter aufweisen, aber auch Items, die von den Versuchspersonen unterschiedlich interpretiert wurden, werden bei dieser Herangehensweise sehr deutlich identifiziert und können im Zuge einer Fragebogenrevision überarbeitet oder ausgetauscht werden.

Die genutzte Simulationsstudie ermöglichte es, verschiedene Konfigurationen für den adaptiven Test auszuprobieren und miteinander zu vergleichen. Bei der Simulation adaptiver Tests können der Testverlauf und weitere Daten erfasst werden. Die Analyse dieser Informationen ermöglicht ein tiefergehendes Verständnis über die Auswirkungen der unterschiedlichen Konfigurationsoptionen. Betrachtet man die anfallenden Item-Nutzungsstatistiken, so wird deutlich, welche Items konfigurationsbedingt häufig oder selten eingesetzt werden. Diese Information konnte bei der Testentwicklung und -überarbeitung genutzt werden, um zum Beispiel gezielt Items zu entwickeln, die in bestimmten Bereichen der latenten Fähigkeitsausprägung besonders gut differenzieren können.

Die Validierung der Ergebnisse der simulierten adaptiven Testung wurde anhand der Befragungsdaten der Skala zur Erfassung von Service-Qualität aus dem ersten Teil der Studie durchgeführt. Diese naheliegende Strategie sollte in künftigen Untersuchungen durch Verhaltensmaße und direkt messbare, zählbare Größen wie Verkaufsstatistiken oder Wartezeiten ergänzt werden.

Um sicherzustellen, dass die adaptive Erfassung von Service-Qualität nicht einseitig Items mit günstigen Item-Parametern aus bestimmten Domänen bzw. Subfacetten des Konstrukts auswählt, können verschiedene Verfahren, die in Kapitel 7.5.4 beschrieben wurden, eingesetzt werden.

8.3 Anwendungsmöglichkeiten und Grenzen

Neben den wissenschaftlichen Ergebnissen dieser Arbeit, die genutzt werden können, um das Verständnis von Service-Qualität, Kundenzufriedenheit und dem Erfolg von Organisationen zu erweitern, wurde ein Fragebogen zur Erfassung von Service-Qualität entwickelt. Dieser linear dargebotene Fragebogen wurde so konzipiert, dass er in unterschiedlichen Organisationen aus verschiedenen Branchen eingesetzt werden kann. Das Instrument ermöglicht es Personen, die guten Einblick in die Service Prozesse einer Organisation haben, diese anhand der vorgegeben Fragen einzuschätzen. Das Ergebnis der Gesamtskala kann für eine allgemeine Einschätzung der Service-Qualität genutzt werden. Um im Detail zu prüfen, in welchen Bereichen des Service-Managements Verbesserungspotenziale diagnostiziert wurden, können die Ergebnisse der Subskalen und auch die Ebene der Einzelitems genutzt werden. Da der entwickelte Fragebogen alle relevanten Bereiche von Service-Qualität umfasst, kann er im Service-Management auch als Checkliste eingesetzt werden. Dazu können Einzelpersonen oder Gruppen die einzelnen Fragebogeninhalte im Hinblick auf eine Organisation diskutieren und ableiten, in welchen Bereichen Veränderungen angestrebt werden sollen. Der Fragebogen und die Checkliste können genutzt werden, um die eigene Organisation einzuschätzen, können als Instrumente aber auch von externen Personen mit hoher Expertise im Bereich Service bearbeitet werden. Der Vergleich der Selbst- und Fremdeinschätzung kann wichtige Hinweise liefern, welche Service-Aspekte von außenstehenden Personen anders wahrgenommen werden.

In der Praxis zeigte sich, dass die Skala zur Erfassung von Service-Qualität aufgrund ihres Umfangs und der damit einher gehenden Bearbeitungsdauer für einige Anwendungszwecke ungeeignet ist. Es wurde beobachtet, dass Personen, die kein besonders hohes Interesse an dem Konstrukt haben, die Beantwortung vorzeitig abbrechen oder die Fragen nicht ernsthaft beantworten.

Wenn in einer Studie mehrere Konstrukte erfasst werden sollen oder Service-Qualität in kurzen Abständen wiederholt gemessen werden soll, bietet sich die adaptive Testform aufgrund ihrer hohen Ökonomie besonders an. Sie ermöglicht es, mit einem Bruchteil der Fragen der Skala zur Erfassung von Service-Qualität bereits eine sehr gute Schätzung der Gesamt-Service-Qualität zu erhalten. Wenn das Konstrukt Service-Qualität möglichst umfassend erfasst werden soll, um zum Beispiel auf der Ebene der Subfacetten oder Einzelitems Verbesserungsansätze zu identifizieren, wird die Nutzung der vollständigen Skala zur Erfassung von Service-Qualität empfohlen, da die gemessene Gesamt-Service-Qualität des adaptiven Tests hierzu bedingt eingesetzt werden kann. Im Sinne einer sequenziellen Diagnostik kann empfohlen werden, zunächst den adaptiven Test einzusetzen, um eine erste Einschätzung der Service-Qualität zu erhalten. Sollten sich dabei unerwartete Ergebnisse zeigen, sollte anschließend die gesamte Skala zur Erfassung von Service-Qualität genutzt werden, um im Detail zu diagnostizieren, wo Verbesserungspotenziale aufgezeigt werden können.

Wenn die Service-Qualität einer Organisation gesteigert werden soll, empfiehlt es sich, die Ergebnisse der Messung von Service-Qualität im Rahmen einer kritischen Interpretation zunächst mit Mitgliedern aus der Organisation zu diskutieren. Für die Planung von wirksamen Interventionen sollten nicht nur die Testergebnisse, sondern auch die qualitativen Einschätzungen der Organisationsmitglieder berücksichtigt werden.

Die Ergebnisse dieser Studie sind in den Kriterienkatalog der Zertifizierung „TÜV SÜD geprüfte ServiceExcellence“ eingeflossen, nach denen sich Unternehmen von der TÜV SÜD Management Service zertifizieren lassen können. Darüber hinaus wurde in einem Arbeitskreis mit weiteren Experten die DIN SPEC 77224 erarbeitet, die ein Managementsystem beschreibt, das „Service Excellence“ genannt wird (DIN SPEC 77224:2011-07).

8.4 Praxiserprobung des adaptiven Tests zur Erfassung von Service-Qualität

Bei der Erprobung des entwickelten adaptiven Tests zur Erfassung von Service-Qualität in der Praxis konnte gezeigt werden, dass die Instruktionen und der Testablauf für die Versuchspersonen verständlich waren, so dass eine selbstständige Bearbeitung ohne weitere Unterstützung möglich ist. Die Teilnehmenden meldeten zurück, dass sie den Test als benutzerfreundlich empfanden, und bewerteten die direkte Ergebnisrückmeldung im Anschluss an die Testung positiv. Das Testergebnis und die generierte Rückmeldungsgrafik konnten problemlos interpretiert werden. Es wurde angeregt, dass branchenspezifische Referenz- bzw. Normwerte bei der Interpretation nützlich wären. Eine Person wünschte sich noch klarere Instruktionen, insbesondere im Blick auf die fehlende Möglichkeit, bereits gegebene Antworten zu überarbeiten.

Ein Test, in dem mehrere Personen gleichzeitig auf das System zugriffen, zeigte, dass gängige Webserver in der Lage sind, das Testsystem zu hosten, und dass die Berechnungen, die im Hintergrund durchgeführt werden müssen, so schnell ablaufen, dass den Anwendern selbst bei mehreren gleichzeitigen Testungen keine merkliche Verzögerung auffiel.

Um die Idee des adaptiven Testens in Organisationen zu verbreiten, wurde das Konzept mehreren Personen aus verschiedenen Organisationen vorgestellt. Dabei wurde die Erfahrung gemacht, dass die Logik adaptiver Tests und die damit mögliche Zeit- und Kostenersparnis leicht überzeugt. Gleichzeitig wurde deutlich, dass Befragungen in Organisationen häufig das Ergebnis langer politischer Abstimmungsprozesse sind und Veränderungen am Befragungsinstrument nur unter Berücksichtigung aller Interessensgruppen möglich sind. Ein adaptiver Test, bei dem ein Algorithmus auswählt, welche Fragen zur Testung einer Einzelperson genutzt werden, wurde vor diesem Hintergrund als kritisch gesehen.

Eine Analyse von in der Praxis eingesetzten Befragungsinstrumente, wie zum Beispiel Mitarbeiterbefragungen, zeigte, dass diese in der Regel bereits sehr kurz gestaltet sind. Meist zielen

diese Instrumente nicht darauf ab, latente Konstrukte umfassend zu erheben, so dass sie durch die Nutzung adaptiver Testverfahren nur geringfügig verkürzt werden können.

8.5 Zukünftige Forschung

Betrachtet man das in Abbildung 9 dargestellte Modell zu den Zusammenhängen innerhalb der Service-Profit Chain von Heskett et al. (1994), so wird deutlich, dass neben den Annahmen, dass Service-Qualität zu Kundenzufriedenheit führt und Kundenzufriedenheit den Erfolg vorhersagen kann, bereits in diesem Modell davon ausgegangen wurde, dass der Erfolg einer Organisation sich auf deren Service-Qualität auswirkt. Damit ist das letzte Glied der Kette, der Erfolg, mit dem ersten Glied, der Service-Qualität, verbunden und es resultiert ein geschlossener Kreislauf. Ein solches Modell könnte als Zirkelschluss interpretiert und deshalb als logisch und wissenschaftlich unzulässig betrachtet werden. Theoretisch ist die Annahme, dass es auch eine Rückwirkung des Erfolgs von Organisationen auf deren Service-Qualität gibt, durchaus begründbar. Solche Modelle stellen die in der psychologischen Forschung gängigen statistischen Methoden, die hier in der Regel nicht ohne weiteres anwendbar sind, vor Herausforderungen. Um die Kausalität und die zeitliche Dynamik solcher Modelle zu prüfen, müssten alle enthaltenen Konstrukte zu mehreren Messzeitpunkten erfasst und autoregressive Modelle genutzt werden (Pakpahan, Hoffmann & Kröger, 2017).

Der Entwicklung des adaptiven Tests im zweiten Teil dieser Studie lag die Annahme zugrunde, dass alle Items der Skala zur Erfassung von Service-Qualität Indikatoren für die Gesamt-Service-Qualität sind. Damit wurde die Struktur der vier Subdimensionen aufgegeben und festgelegt, dass der adaptive Test nur ein Testergebnis für die Gesamt-Service-Qualität einer Organisation diagnostiziert. Wenn der adaptive Test auch Schätzungen für die vier Subdimensionen bereitstellen soll, könnte er als multidimensionaler adaptiver Test weiterentwickelt werden (Yao, Pommerich & Segall, 2014). Grundlage dafür wäre die Bestimmung der Item-Parameter mit dem Ansatz der multidimensionalen Item Response Theory (McDonald, 2000).

Der entwickelte adaptive Test zur Erfassung von Service-Qualität könnte mit überschaubarem Aufwand erweitert werden, so dass auch weitere Metadaten, wie zum Beispiel Reaktions- bzw. Bearbeitungszeiten, erfasst werden. Dies bietet die Möglichkeit, in weiteren Studien zu untersuchen, ob es Zusammenhänge zwischen diesen Metadaten und dem Antwortverhalten von Versuchspersonen gibt.

Da der entwickelte adaptive Test zur Erfassung von Service-Qualität auf gängigen Internettechnologien basiert, könnten auch Item-Formate, die zum Beispiel Video- oder Bildmaterial enthalten, integriert werden. Wenn künftige Studien herausfinden, welche Elemente der Items genau für ihre psychometrischen Eigenschaften verantwortlich sind, könnte das adaptive Testsystem auch gezielt Items generieren. Wenn sichergestellt ist, dass das System ausschließlich

mit ernsthaften Antwortdaten in Kontakt kommt, könnten diese genutzt werden, um die Item-Parameter in regelmäßigen Abständen neu zu kalibrieren.

In weiteren Studien mit adaptiven Tests sollte der Möglichkeit, verschiedene Variablen zu komplexeren Abbruchkriterien zu verknüpfen, die in Kapitel 7.4.4 beschrieben wurden, Aufmerksamkeit geschenkt werden. Es wäre beispielsweise denkbar, dass das adaptive Testsystem nicht ernsthaftes Antwortverhalten erkennt, dies an die Versuchsperson zurückmeldet und die Testung bei weiteren unrealistischen Antworten abbricht.

8.6 Ausblick

Wie sich in dieser Studie gezeigt hat, steht Service-Qualität im Zusammenhang mit Kundenzufriedenheit und dem Erfolg von Organisationen und ist damit eine wichtige Kenngröße. Organisationen, die langfristig Wert auf zufriedene Kundschaft und wirtschaftlichen Erfolg legen, sollten ihre Service-Qualität deshalb im Blick haben. Um die Qualität von Dienstleistungen sicherzustellen, benötigen Organisationen, die ausgezeichnete Service-Qualität bieten möchten, zuverlässige und ökonomisch Messverfahren für das Service-Management. Die in dieser Arbeit entwickelte Skala zur Erfassung von Service-Qualität und der entwickelte adaptive Test bieten zwei Optionen, um Service-Qualität unter verschiedenen Rahmenbedingungen zu erfassen. Wird Service-Qualität regelmäßig erfasst, so stellt diese Messung in gewisser Weise auch eine Form von Intervention dar. Die regelmäßige Auseinandersetzung mit relevanten Fragen aus dem Bereich Service sorgt dafür, dass das Thema salient bleibt und sich auf das tägliche Handeln auswirkt.

„Developing a reliable instrument is a journey rather than a destination" – Diese Aussage trifft auch auf die Weiterentwicklung der Skala zur Erfassung von Service-Qualität zu (Brown, Gummesson, Edvardsson & Gustavsson, 1991, S. 17). Durch die intensive Auseinandersetzung mit den psychometrischen Eigenschaften der Items und deren Nützlichkeit bei der adaptiven Testung von Service-Qualität wurde deutlich, dass die Skala um Items mit sehr hohen und sehr niedrigen Item-Schwierigkeiten erweitert werden könnte, so dass auch in den Bereichen geringer Service-Qualität und exzellenten Services ausreichend gut differenziert gemessen werden kann. Ein etwas breiter aufgestellter Item-Pool würde auch für den adaptiven Test sehr nützlich sein und ermöglichen, auch in den Bereichen sehr geringer und sehr hoher Service-Qualität ökonomisch und präzise zu diagnostizieren.

Die Diagnostik psychologischer Konstrukte hat sich in den letzten Jahren weiterentwickelt und professionalisiert. Durch die Digitalisierung und die zunehmende Nutzung des Internets haben sich Möglichkeiten für Onlinebefragungen ergeben, die klassische Papier-und-Stift-Verfahren immer weiter verdrängen. Moderne Onlinemedien sind von kurzen, übersichtlichen Inhalten geprägt. Das Nutzungsverhalten ist selektiv und ausführlich; Inhalte werden häufig

nur überblicksartig verarbeitet. Für die Erfassung von psychologischen Konstrukten, die bislang normalerweise das Beantworten von vielen, sehr ähnlichen Fragebogenitems umfasst, stellt dieses neue Umfeld mit seinen Nutzungsgewohnheiten eine Herausforderung dar. Vor diesem Hintergrund ist der Einsatz adaptiver Tests eine zukunftsrelevante und elegante Lösung, die reliable und valide Messungen ermöglicht, ohne dabei auf langwierige Befragungsinstrumente angewiesen zu sein.

Literaturverzeichnis

Ackermann, M. P. (1992). Rahmenbedingungen für die Effizienz von Kleingruppenaktivitäten. In W. Bungard (Hrsg.), *Qualitätszirkel in der Arbeitswelt: Ziele, Erfahrungen, Probleme. Beiträge zur Organisationspsychologie*. Göttingen: Verlag für Angewandte Psychologie.

Adair, J. G. (1984). The Hawthorne effect: A reconsideration of the methodological artifact. *Journal of Applied Psychology*, *69*(2), 334.

Adams, J. S. (1963). Towards an understanding of inequity. *The Journal of Abnormal and Social Psychology*, *67*(5), 422. https://doi.org/10.1037/h0040968

Ali, M. & Raza, S. A. (2017). Service quality perception and customer satisfaction in Islamic banks of Pakistan. The modified SERVQUAL model. *Total Quality Management & Business Excellence*, *28*(5-6), 559–577. https://doi.org/10.1080/14783363.2015.1100517

Anderson, E. W. & Sullivan, M. W. (1993). The Antecedents and Consequences of Customer Satisfaction for Firms. *Marketing science*, *12*(2), 125–143.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561–573. https://doi.org/10.1007/BF02293814

Ashkanasy, N. M., Wilderom, C. & Peterson, M. F. (Hrsg.). (2000). *Handbook of Organizational Culture and Climate*. Thousand Oaks, Calif: Sage Publications.

Auh, S., Menguc, B., Fisher, M. & Haddad, A. (2011). The Contingency Effect of Service Employee Personalities on Service Climate: Getting Employee Perceptions Aligned Can Reduce Personality Effects. *Journal of Service Research*, *14*(4), 426–441. https://doi.org/10.1177/1094670511421521

Baker, F. B. (1992). *Item Response Theory: Parameter Estimation Techniques*. New York: M. Dekker.

Barrada, J. R., Olea, J., Ponsoda, V. & Abad, F. J. (2009). Item Selection Rules in Computerized Adaptive Testing. *Methodology*, *5*(1), 7–17. https://doi.org/10.1027/1614-2241.5.1.7

Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences*, *42*(5), 815–824. https://doi.org/10.1016/j.paid.2006.09.018

Barros Jerônimo, T. de & Medeiros, D. (2014). Measuring quality service. *International Journal of Quality & Reliability Management*, *31*(6), 652–664. https://doi.org/10.1108/IJQRM-06-2012-0095

Beauducel, A. & Herzberg, P. Y. (2006). On the Performance of Maximum Likelihood Versus Means and Variance Adjusted Weighted Least Squares Estimation in CFA. *Structural Equation Modeling: A Multidisciplinary Journal*, *13*(2), 186–203. https://doi.org/10.1207/s15328007sem1302_2

Benson, N., Hulac, D. M. & Kranzler, J. H. (2010). Independent examination of the Wechsler Adult Intelligence Scale—Fourth Edition (WAIS-IV): What does the WAIS-IV measure? *Psychological Assessment*, *22*(1), 121–130. https://doi.org/10.1037/a0017767

Berry, L. L. & Parasuraman, A. (1992). *Service-Marketing*. Frankfurt/Main, New York: Campus-Verlag.

Bettencourt, L. A., Gwinner, K. P. & Meuter, M. L. (2001). A Comparison of Attitude, Personality, and Knowledge Predictors of Service-Oriented Organizational Citizenship Behaviors. *Journal of Applied Psychology*, *86*(1), 29–41. https://doi.org/10.1037//0021-9010.86.1.29

Betz, N. E. & Weiss, D. J. (1976). *Psychological effects of immediate knowledge of results and adaptive ability testing. (Research Report 76–4).* Minneapolis, MN: University of Minnesota, Department of Psychology.

Bitner, M. J. & Hubbert, A. R. (1994). Encounter Satisfaction Versus Overall Satisfaction Versus Quality. In R. T. Rust (ed.), *Service quality. New Directions in Theory and Practice* (S. 72–94). Thousand Oaks, California: Sage Publications.

Bock, R. D. & Mislevy, R. J. (1982). Adaptive EAP Estimation of Ability in a Microcomputer Environment. *Applied Psychological Measurement*, *6*(4), 431–444. https://doi.org/10.1177/014662168200600405

Borucki, C. C. & Burke, M. J. (1999). An examination of service-related antecedents to retail store performance. *Journal of Organizational Behavior*, *20*(6), 943–962. https://doi.org/10.1002/(SICI)1099-1379(199911)20:6<943::AID-JOB976>3.0.CO;2-9

Bovermann, A. (2013). *Dienstleistungsqualität durch Total Quality Management*. Wiesbaden: Springer.

Bowen, D. E. & Schneider, B. (2014). A Service Climate Synthesis and Future Research Agenda. *Journal of Service Research*, *17*(1), 5–22. https://doi.org/10.1177/1094670513491633

Brady, M. K. & Cronin Jr, J. J. (2001). Customer Orientation: Effects on Customer Service Perceptions and Outcome Behaviors. *Journal of Service Research*, *3*(3), 241–251. https://doi.org/10.1177/109467050133005

Brady, M. K., Cronin Jr, J. J. & Cronin, J. J. (2001). Some New Thoughts on Conceptualizing Perceived Service Quality: A Hierarchical Approach. *Journal of Marketing*, *65*(3), 34–49. https://doi.org/10.1509/jmkg.65.3.34.18334

Brady, M. K. & Robertson, C. J. (2001). Searching for a Consensus on the Antecedent Role of Service Quality and Satisfaction: An Exploratory Cross-National Study. *Journal of Business research*, *51*(1), 53–60. https://doi.org/10.1016/S0148-2963(99)00041-7

Brauer, J.-P. & Kühme, E. U. (1996). *DIN EN ISO 9000-9004. Gestaltungshilfen zum Aufbau ihres Qualitätsmanagementsystems*. München, Wien: Hanser.

Braun, O. L. & Haferburg, M. (2001). Kundenzufriedenheit: Theorie, Messung und Intervention. In O. L. Braun, J. Abendschein, M. Haferburg & S. Mihailovic (Hrsg.), *Kundenzufriedenheit und psychologisches Qualitätsmanagement* (S. 13–27). Heidelberg: Editon GP.

Braun, O. L. & Müssigmann, M. J. (2009a). Einleitung, Theorie und Ablauf einer Kundenbefragung. In O. L. Braun & M. J. Müssigmann (Hrsg.), *Kundenbefragungen Eine Sammlung von Fallstudien* (S. 7–22). Lengerich: Pabst Science Publishers.

Braun, O. L. & Müssigmann, M. J. (Hrsg.). (2009b). *Kundenbefragungen Eine Sammlung von Fallstudien*. Lengerich: Pabst Science Publishers.

Braun, T. & Koch, J. (2002). Qualitätsmanagement: Entwicklung, Problemfelder und ein integrativer Lösungsvorschlag. *Organisationsberatung, Supervision, Coaching*, *9*(2), 149–159. https://doi.org/10.1007/s11613-002-0015-x

Brown, S. W., Gummesson, E., Edvardsson, B. & Gustavsson, B. (Hrsg.). (1991). *Service Quality: Multidisciplinary and Multinational Perspectives*. Lexington, Mass: Lexington Books.

Brown, T. J., Churchill Jr, G. A. & Peter, J. P. (1993). Improving the Measurement of Service Quality. *Journal of Retailing*, *69*(1), 127–139. https://doi.org/10.1016/S0022-4359(05)80006-5

Brown, T. J., Mowen, J. C., Donavan, D. T. & Licata, J. W. (2002). The Customer Orientation of Service Workers: Personality Trait Effects on Self-and Supervisor Performance Ratings. *Journal of Marketing Research*, *39*(1), 110–119. https://doi.org/10.1509/jmkr.39.1.110.18928

Bruhn, M. (2010). *Qualitätsmanagement für Dienstleistungen: Grundlagen, Konzepte, Methoden* (8., überarb. u. erw. Aufl.). Springer Berlin Heidelberg.

Bungard, W. (1991). *Qualitätszirkel – Ein soziotechnisches Instrument auf dem Prüfstand* (Mannheimer Schriften zur Arbeits- und Organisationspsychologie, Bd. 1). Ludwigshafen/Rh.: Ehrenhof.

Bungard, W. (Hrsg.). (1992). *Qualitätszirkel in der Arbeitswelt: Ziele, Erfahrungen, Probleme. Beiträge zur Organisationspsychologie*. Göttingen: Verlag für Angewandte Psychologie.

Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. https://doi.org/10.1037/h0046016

Canivez, G. L. (2013). Incremental criterion validity of WAIS–IV factor index scores: Relationships with WIAT–II and WIAT–III subtest and composite scores. *Psychological Assessment*, *25*(2), 484–495. https://doi.org/10.1037/a0032092

Cardozo, R. N. (1965). An Experimental Study of Customer Effort, Expectation, and Satisfaction. *Journal of Marketing Research*, *2*(3), 244–249. https://doi.org/10.2307/3150182

Carroll, B. J. (2013). Computerized Adaptive Test–Depression Inventory Not Ready for Prime Time. *JAMA Psychiatry*, *70*(7), 763–765. https://doi.org/10.1001/jamapsychiatry.2013.1318

Central Intelligence Agency (CIA, Hrsg.). (2020). *The World Factbook. GDP – Composition by Sectors of origin*. Zugriff am 17.10.2020. Verfügbar unter https://www.cia.gov/library/publications/the-world-factbook/fields/214.html#GM

Chang, S.-W. & Ansley, T. N. (2003). A Comparative Study of Item Exposure Control Methods in Computerized Adaptive Testing. *Journal of Educational Measurement*, *40*(1), 71–103. https://doi.org/10.1111/j.1745-3984.2003.tb01097.x

Chen, Z., Zhu, J. & Zhou, M. (2015). How Does a Servant Leader Fuel the Service Fire? A Multilevel Model of Servant Leadership, Individual Self Identity, Group Competition Climate, and Customer Service Performance. *Journal of Applied Psychology*, *100*(2), 511–521. https://doi.org/10.1037/a0038036

Cheng, Y., Chang, H.-H. & Yi, Q. (2007). Two-Phase Item Selection Procedure for Flexible Content Balancing in CAT. *Applied Psychological Measurement*, *31*(6), 467–482. https://doi.org/10.1177/0146621606292933

Cheung, G. W. (2008). Testing Equivalence in the Structure, Means, and Variances of Higher-Order Constructs With Structural Equation Modeling. *Organizational Research Methods*, *11*(3), 593–613. https://doi.org/10.1177/1094428106298973

Chien, T.-W., Wang, W.-C., Huang, S.-Y., Lai, W.-P. & Chow, J. C. (2011). A web-based computerized adaptive testing (CAT) to assess patient perception in hospitalization. *Journal of Medical Internet Research*, *13*(3), 1–10. https://doi.org/10.2196/jmir.1785

Choi, S. W. (2009). Firestar. Computerized Adaptive Testing Simulation Program for Polytomous Item Response Theory Models. *Applied Psychological Measurement*, *33*(8), 644–645. https://doi.org/10.1177/0146621608329892

Choi, S. W., Grady, M. W. & Dodd, B. G. (2010). A New Stopping Rule for Computerized Adaptive Testing. *Educational and Psychological Measurement*, *70*(6), 1–17. https://doi.org/10.1177/0013164410387338

Choi, S. W., Podrabsky, T. & McKinney, N. (2012). Firestar- D. *Applied Psychological Measurement*, *36*(1), 67–68. https://doi.org/10.1177/0146621611406107

Choi, S. W., Reise, S. P., Pilkonis, P. A., Hays, R. D. & Cella, D. (2010). Efficiency of static and computer adaptive short forms compared to full-length measures of depressive symptoms. *Quality of Life Research*, *19*(1), 125–136. https://doi.org/10.1007/s11136-009-9560-5

Choi, S. W. & Swartz, R. J. (2009). Comparison of CAT Item Selection Criteria for Polytomous Items. *Applied Psychological Measurement*, *33*(6), 419–440. https://doi.org/10.1177/0146621608327801

Colquitt, J. A. (2001). On the Dimensionality of Organizational Justice: A Construct Validation of a Measure. *The Journal of Applied Psychology*, *86*(3), 386–400. https://doi.org/10.1037//0021-9010.86.3.386

Colquitt, J. A., Conlon, D. E., Wesson, M. J., Porter, C. O. & Ng, K. Y. (2001). Justice at the Millennium: A Meta-Analytic Review of 25 years of Organizational ustice Research. *The Journal of Applied Psychology*, *86*(3), 425–445. https://doi.org/10.1037//0021-9010.86.3.425

Cran, D. J. (1994). Towards Validation of the Service Orientation Construct. *The Service Industries Journal*, *14*(1), 34–44. https://doi.org/10.1080/02642069400000003

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

Cronin Jr, J. J., Brady, M. K. & Hult, G. T. M. (2000). Assessing the Effects of Quality, Value, and Customer Satisfaction on Consumer Behavioral Intentions in Service Environments. *Journal of Retailing*, *76*(2), 193–218. https://doi.org/10.1016/S0022-4359(00)00028-2

Cronin Jr, J. J. & Taylor, S. A. (1992). Measuring Service Quality: A Reexamination and Extension. *Journal of Marketing*, *56*(3), 55–68. https://doi.org/10.2307/1252296

Cronin Jr, J. J. & Taylor, S. A. (1994). SERVPERF versus SERVQUAL: Reconciling Performance-Based and Perceptions-Minus-Expectations Measurement of Service Quality. *Journal of Marketing*, *58*(1), 125–131. https://doi.org/10.1177/002224299405800110

Cropanzano, R. & Greenberg, J. (1997). Progress in Organizational Justice: Tunneling Through the Maze. In C. L. Cooper & I. T. Robertson (Hrsg.), *International Review of Industrial and Organizational Psychology* (Bd. 12, S. 317–372).

Csikszentmihalyi, M. (2014). *Flow. Das Geheimnis des Glücks* (17. Auflage). Stuttgart: Klett-Cotta.

Dabholkar, P. A., Shepherd, C. D. & Thorpe, D. I. (2000). A Comprehensive Framework for Fervice Quality: An Investigation of Critical Conceptual and Measurement Issues Through a Longitudinal Study. *Journal of Retailing*, *76*(2), 139–173. https://doi.org/10.1016/S0022-4359(00)00029-4

Dagger, T. S. & Sweeney, J. C. (2007). Service Quality Attribute Weights: How Do Novice and Longer-Term Customers Construct Service Quality Perceptions? *Journal of Service Research*, *10*(1), 22–42. https://doi.org/10.1177/1094670507303010

Dale, A. & Wooler, S. (1991). Strategy and Organization for Service. In S. W. Brown, E. Gummesson, B. Edvardsson & B. Gustavsson (Hrsg.), *Service Quality: Multidisciplinary and Multinational Perspectives* (S. 191–204). Lexington, Mass: Lexington Books.

De Beurs, D. P. de, de Vries, A. L. de, de Groot, M. H. de, de Keijser, J. de & Kerkhof, A. J. (2014). Applying Computer Adaptive Testing to Optimize Online Assessment of Suicidal Behavior: A Simulation Study. *Journal of Medical Internet Research*, *16*(9), 1–11. https://doi.org/10.2196/jmir.3511

Dean, A. M. (2004). Links between organisational and customer variables in service delivery: Evidence, contradictions and challenges. *International Journal of Service Industry Management*, *15*(4), 332–350. https://doi.org/10.1108/09564230410552031

DIN SPEC, 77224:2011-07 (2011). *Erzielung von Kundenbegeisterung durch Service Excellence*.

DIN EN ISO 9000:2015-11 (2015). *Qualitätsmanagementsysteme – Grundlagen und Begriffe*.

Deville, C. (1993). Flow as a Testing Ideal. *Rasch Measurement Transactions*, *7*(3), 308.

Dienhart, J. R., Gregoire, M. B., Downey, R. G. & Knight, P. K. (1992). Service orientation of restaurant employees. *Journal of Hospitality & Tourism Research*, *11*(4), 331–346. https://doi.org/10.1016/0278-4319(92)90050-6

Dietz, J., Pugh, S. D. & Wiley, J. W. (2004). Climate Effects on Customer Attitudes: An Examination of Boundary Conditions. *Academy of Management Journal*, *47*(1), 81–92. https://doi.org/10.2307/20159561

DiStefano, C. & Morgan, G. B. (2014). A Comparison of Diagonal Weighted Least Squares Robust Estimation Techniques for Ordinal Data. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(3), 425–438. https://doi.org/10.1080/10705511.2014.915373

Donavan, D. T., Brown, T. J. & Mowen, J. C. (2004). Internal Benefits of Service-Worker Customer Orientation: Job Satisfaction, Commitment, and Organizational Citizenship Behaviors. *Journal of Marketing*, *68*(1), 128–146. https://doi.org/10.1509/jmkg.68.1.128.24034

Dunn, T. J., Baguley, T. & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, *105*(3), 399–412. https://doi.org/10.1111/bjop.12046

Ehrhart, K. H., Witt, L. A., Schneider, B. & Perry, S. J. (2011). Service Employees Give as They Get: Internal Service as a Moderator of the Service Climate-Service Outcomes Link. *Journal of Applied Psychology*, *96*(2), 423–431. https://doi.org/10.1037/a0022071

Eid, M. & Diener, E. (Hrsg.). (2006). *Handbook of multimethod measurement in psychology* (1. ed.). Washington, DC: American Psychological Association.

Eisingerich, A. B. & Bell, S. J. (2008). Perceived Service Quality and Customer Trust: Does Enhancing Customers' Service Knowledge Matter? *Journal of Service Research*, *10*(3), 256–268. https://doi.org/10.1177/1094670507310769

Embretson, S. E. & Reise, S. P. (2000). *Item Response Theory for Psychologists* (Multivariate applications book series). Mahwah, N.J: L. Erlbaum Associates.

Epskamp, S., Cramer, A. O. J., Waldorp, L. J., Schmittmann, V. D. & Borsboom, D. (2012). qgraph: Network Visualizations of Relationships in Psychometric Data // qgraph : Network Visualizations of Relationships in Psychometric Data. *Journal of Statistical Software*, *48*(4), 1–18. https://doi.org/10.18637/jss.v048.i04

Epstein, R. (1984). The Principle of Parsimony and Some Applications in Psychology. *Journal of Mind and Behavior*, *5*(2), 119–130.

Etemad-Sajadi, R., Way, S. A. & Bohrer, L. (2016). Airline Passenger Loyalty. The Distinct Effects of Airline Passenger Perceived Pre-Flight and In-Flight Service Quality. *Cornell Hospitality Quarterly*, *57*(2), 219–225. https://doi.org/10.1177/1938965516630622

Fassnacht, M. & Koese, I. (2006). Quality of Electronic Services. Conceptualizing and Testing a Hierarchical Model. *Journal of Service Research*, *9*(1), 19–37. https://doi.org/10.1177/1094670506289531

Finger, M. S. & Ones, D. S. (1999). Psychometric equivalence of the computer and booklet forms of the MMPI: A meta-analysis. *Psychological Assessment*, *11*(1), 58–66. https://doi.org/10.1037/1040-3590.11.1.58

Finn, A. & Kayandé, U. (1999). Unmasking a Phantom: A Psychometric Assessment of Mystery Shopping. *Journal of Retailing*, *75*(2), 195–217. https://doi.org/10.1016/S0022-4359(99)00004-4

Fischer, C., Braun, O. L., Kehr, F. & Schreiber, W. H. (2012). Zur Beziehung zwischen Dienstleistungsqualität, Image und Kundenzufriedenheit am Beispiel einer Polizeibehörde. *Polizei & Wissenschaft*, *1*, 2–12.

Fliege, H., Becker, J., Walter, O. B., Bjorner, J. B., Klapp, B. F. & Rose, M. (2005). Development of a Computer-adaptive Test for Depression (D-CAT). *Quality of Life Research*, *14*(10), 2277–2291. https://doi.org/10.1007/s11136-005-6651-9

Franke, G. H. (1997). "The Whole is More than the Sum of its Parts": The Effects of Grouping and Randomizing Items on the Reliability and Validity of Questionnaires. *European Journal of Psychological Assessment*, *13*(2), 67–74. https://doi.org/10.1027/1015-5759.13.2.67

Frey, A., Hartig, J. & Moosbrugger, H. (2009). Effekte des adaptiven Testens auf die Motivation zur Testbearbeitung am Beispiel des Frankfurter Adaptiven Konzentrationsleistungs-Tests. *Diagnostica*, *55*(1), 20–28. https://doi.org/10.1026/0012-1924.55.1.20

Frey, D. (1997). König Kunde als deutscher Störfaktor – Unternehmer erreichen mehr Servicementalität durch die fünf Prinzipien Kennen, Können, Wollen, Sollen und Dürfen. *Süddeutsche Zeitung*.

Fruchterman, T. M. J. & Reingold, E. M. (1991). Graph Drawing by Force-directed Placement. *Software: Practice and experience*, *21*(11), 1129–1164. https://doi.org/10.1002/spe.4380211102

Fürst, A. (2011). Verfahren zur Messung der Kundenzufriedenheit im Überblick. In C. Homburg (Hrsg.), *Kundenzufriedenheit: Konzepte – Methoden – Erfahrungen* (S. 124–153). Wiesbaden: Gabler Verlag.

Geerlings, H., van der Linden, Wim J. & Glas, C. A. W. (2013). Optimal Test Design With Rule-Based Item Generation. *Applied Psychological Measurement*, *37*(2), 140–161. https://doi.org/10.1177/0146621612468313

Gelade, G. A. & Young, S. (2005). Test of a service profit chain model in the retail banking sector. *Journal of Occupational and Organizational Psychology*, *78*(1), 1–22. https://doi.org/10.1348/096317904X22926

Gibbons, R. D., Weiss, D. J., Pilkonis, P. A., Frank, E., Moore, T., Kim, J. B. et al. (2012). Development of a Computerized Adaptive Test for Depression. *Archives of General Psychiatry*, *69*(11), 1104–1112. https://doi.org/10.1001/archgenpsychiatry.2012.14

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587–606. https://doi.org/10.1016/j.socec.2004.09.033

Golder, P. N., Mitra, D. & Moorman, C. (2012). What Is Quality? An Integrative Framework of Processes and States. *Journal of Marketing*, *76*(4), 1–23. https://doi.org/10.2307/41714496

Gorin, J. S., Dodd, B. G., Fitzpatrick, S. J. & Shieh, Y. Y. (2005). Computerized Adaptive Testing With the Partial Credit Model: Estimation Procedures, Population Distributions, and Item Pool Characteristics. *Applied Psychological Measurement*, *29*(6), 433–456. https://doi.org/10.1177/0146621605280072

Gounaris, S. P., Stathakopoulos, V. & Athanassopoulos, A. D. (2003). Antecedents to perceived service quality. An exploratory study in the banking industry. *International Journal of Bank Marketing*, *21*(4), 168–190. https://doi.org/10.1108/02652320310479178

Grönroos, C. (1984). A Service Quality Model and its Marketing Implications. *European Journal of Marketing*, *18*(4), 36–44. https://doi.org/10.1108/EUM0000000004784

Groves, J., Gregoire, M. B. & Downey, R. G. (1995). Relationship Between the Service Orientation of Employees and Operational Indicators in a Multiunit Restaurant Corporation. *Journal of Hospitality & Tourism Research*, *19*(3), 33–43. https://doi.org/10.1177/109634809501900305

Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1995). *Fundamentals of Item Response Theory* (Measurement methods for the social sciences, vol. 2, 4. [ed.]. Newbury Park: Sage Publ.

Han, K. C. T. (2013). Item Pocket Method to Allow Response Review and Change in Computerized Adaptive Testing. *Applied Psychological Measurement*, *37*(4), 259–275. https://doi.org/10.1177/0146621612473638

Han, K. C. T. (2016). Maximum Likelihood Score Estimation Method with Fences for Short-Length Tests and Computerized Adaptive Tests. *Applied Psychological Measurement*, *40*(4), 289–301. https://doi.org/10.1177/0146621616631317

Han, K. C. T. (2018). Components of the item selection algorithm in computerized adaptive testing. *Journal of Educational Evaluation for Health Professions*, *15*, 1–13. https://doi.org/10.3352/jeehp.2018.15.7

Harris, E. G., Mowen, J. C. & Brown, T. J. (2005). Re-examining Salesperson Goal Orientations: Personality Influencers, Customer Orientation, and Work Satisfaction. *Journal of the Academy of Marketing Science*, *33*(1), 19–35. https://doi.org/10.1177/0092070304267927

Harrison-Walker, L. J. (2016). The Measurement of Word-of-Mouth Communication and an Investigation of Service Quality and Customer Commitment aAs Potential Antecedents. *Journal of Service Research*, *4*(1), 60–75. https://doi.org/10.1177/109467050141006

Hartline, M. D. & Ferrell, O. C. (1996). The Management of Customer-Contact Service Employees. An Empirical Investigation. *Journal of Marketing*, *60*(4), 52–70. https://doi.org/10.2307/1251901

Häussermann, H. (1995). *Dienstleistungsgesellschaften* (Edition Suhrkamp, 2. Aufl.). Frankfurt am Main: Suhrkamp.

Heflin, C., Sandberg, J. & Rafail, P. (2009). The Structure of Material Hardship in U.S. Households: An Examination of the Coherence behind Common Measures of Well-Being. *Social Problems*, *56*(4), 746–764. https://doi.org/10.1525/sp.2009.56.4.746

Hegner, F. (1994). Zusammenhänge zwischen „Lean Production", „Kaizen" und „Totalem Qualitätsmanagement". *Arbeit*, *4*(3), 299–319.

Henson, R., Roussos, L., Douglas, J. & He, X. (2008). Cognitive Diagnostic Attribute-Level Discrimination Indices. *Applied Psychological Measurement*, *32*(4), 275–288. https://doi.org/10.1177/0146621607302478

Heskett, J. L. (2014). Notes from the Search for Deep Indicators in Services. *Journal of Service Management*, *25*(3), 298–309. https://doi.org/10.1108/JOSM-04-2014-0105

Heskett, J. L., Jones, T. O., Loveman, G. W., Sasser, W. E. & Schlesinger, L. A. (1994). Putting the Service-Profit Chain to Work. *Harvard business review*, *72*(2), 164–174.

Heskett, J. L., Sasser, W. E. & Hart, C. W. L. (1991). *Bahnbrechender Service: Standards für den Wettbewerb von Morgen*. Frankfurt/Main, New York: Campus-Verlag.

Heskett, J. L., Sasser, W. E. & Schlesinger, L. A. (1997). *The service profit chain. How leading companies link profit and growth to loyalty, satisfaction, and value*. New York, NY: Free Press.

Heskett, J. L., Sasser, W. E. & Schlesinger, L. A. (2003). *The Value Profit Chain. Treat employees like customers and customers like employees*. New York: The Free Press.

Heskett, J. L., Sasser, W. E. & Schlesinger, L. A. (2015). *What Great Service Leaders Know and Do. Creating Breakthroughs in Service Firms*. Oakland, United States: Berrett-Koehler Publishers.

Heskett, J. L., Sasser, W. E. & Wheeler, J. (2010). The Ownership Quotient: Putting the Service Profit Chain to Work for Unbeatable Competitive Advantage. *Journal of Service Management*, *21*(3), 413–417. https://doi.org/10.1108/09564231011050823

Hinrichs, S. (2005). *Qualitätskenntnis. Psychologische Aspekte der Qualität von Arbeit und Produkt*. Lengerich: Pabst Science Publ.

Hogan, J., Hogan, R. & Busch, C. M. (1984). How to measure service orientation. *Journal of Applied Psychology*, *69*(1), 167–173. https://doi.org/10.1037/0021-9010.69.1.167

Hogreve, J., Iseke, A., Derfuss, K. & Eller, T. (2017). The Service–Profit Chain. A Meta-Analytic Test of a Comprehensive Theoretical Framework. *Journal of Marketing*, *81*(3), 41–61. https://doi.org/10.1509/jm.15.0395

Homburg, C. (Hrsg.). (2011). *Kundenzufriedenheit: Konzepte – Methoden – Erfahrungen*. Wiesbaden: Gabler Verlag.

Homburg, C., Müller, M. & Klarmann, M. (2011). When does salespeople's customer orientation lead to customer loyalty? The differential effects of relational and functional customer orientation. *Journal of the Academy of Marketing Science*, *39*(6), 795–812. https://doi.org/10.1007/s11747-010-0220-7

Homburg, C., Wieseke, J. & Hoyer, W. D. (2009). Social Identity and the Service–Profit Chain. *Journal of Marketing*, *73*(2), 38–54. https://doi.org/10.1509/jmkg.73.2.38

Hong, Y., Liao, H., Hu, J. & Jiang, K. (2013). Missing Link in the Service Profit Chain: A Meta-Analytic Review of the Antecedents, Consequences, and Moderators of Service Climate. *Journal of Applied Psychology*, *98*(2), 237–267. https://doi.org/10.1037/a0031666

Hu, L.-t. & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*(4), 424–453. https://doi.org/10.1037//1082-989X.3.4.424

Hu, L.-t. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hui, C. H., Chiu, W. C. K., Yu, P. L. H., Cheng, K. & Tse, H. H. M. (2007). The effects of service climate and the effective leadership behaviour of supervisors on frontline employee service quality: A multi-level analysis. *Journal of Occupational and Organizational Psychology*, *80*(1), 151–172. https://doi.org/10.1348/096317905X89391

Hutchinson, S. R. & Olmos, A. (1998). Behavior of descriptive fit indexes in confirmatory factor analysis using ordered categorical data. *Structural Equation Modeling: A Multidisciplinary Journal*, *5*(4), 344–364. https://doi.org/10.1080/10705519809540111

Irvine, S. H. (2002). The foundations of item generation for mass testing. In S. H. Irvine & P. C. Kyllonen (Hrsg.), *Item generation for test development* (S. 3–34). Mahwah, NJ: Lawrence Erlbaum.

Irvine, S. H. & Kyllonen, P. C. (Hrsg.). (2002). *Item generation for test development*. Mahwah, NJ: Lawrence Erlbaum.

Jackson, D. L., Gillaspy, J. A. & Purc-Stephenson, R. (2009). Reporting Practices in Confirmatory Factor Analysis: An Overview and Some Recommendations. *Psychological Methods*, *14*(1), 6–23. https://doi.org/10.1037/a0014694

Jacobs, C., Heubrock, D. & Petermann, F. (2003). Testinformation. *Diagnostica*, *49*(4), 184–188. https://doi.org/10.1026//0012-1924.49.4.184

Jain, S. K. & Gupta, G. (2004). Measuring Service Quality. Servqual vs. Servperf Scales. *Vikalpa*, *29*(2), 25–38. https://doi.org/10.1177/0256090920040203

Jong, A. d., Ruyter, K. d. & Lemmink, J. (2004). Antecedents and Consequences of the Service Climate in Boundary-Spanning Self-Managing Service Teams. *Journal of Marketing*, *68*(2), 18–35. https://doi.org/10.1509/jmkg.68.2.18.27790

Kamakura, W. A., Mittal, V., Rosa, F. de & Mazzon, J. A. (2002). Assessing the Service-Profit Chain. *Marketing science*, *21*(3), 294–317. https://doi.org/10.1287/mksc.21.3.294.140

Kanning, U. P. & Bergmann, N. (2009). Predictors of customer satisfaction: testing the classical paradigms. *Managing Service Quality: an International Journal*, *19*(4), 377–390. https://doi.org/10.1108/09604520910971511

Kaplan, D. E. (2000). *Structural equation modeling. Foundations and extensions* (vol. 10). Thousand Oaks: Sage Publications.

Kaplan, M., La Torre, J. de & Barrada, J. R. (2015). New Item Selection Methods for Cognitive Diagnosis Computerized Adaptive Testing. *Applied Psychological Measurement*, *39*(3), 167–188. https://doi.org/10.1177/0146621614554650

Keller, L. A. (2000). *Ability Estimation Procedures in Computerized Adaptive Testing*. USA: American Institute of Certified Public Accountants-AICPA Research Concortium-Examination Teams.

Kingsbury, G. (1996). *Item review and adaptive testing* (Annual Meeting of the National Council on Measurement in, Hrsg.). New York.

Kirisci, L., Tarter, R., Reynolds, M., Ridenour, T., Stone, C. & Vanyukov, M. (2012). Computer adaptive testing of liability to addiction: Identifying individuals at risk. *Drug and Alcohol Dependence*, *1231*(1), 79–86. https://doi.org/10.1016/j.drugalcdep.2012.01.016

Klarmann, M. (2011). Die Vergleichbarkeit der Messung als Herausforderung bei internationalen Kundenzufriedenheitsuntersuchungen. In C. Homburg (Hrsg.), *Kundenzufriedenheit: Konzepte – Methoden – Erfahrungen* (S. 228–246). Wiesbaden: Gabler Verlag.

Klodt, H., Maurer, R. & Schimmelpfennig, A. (1997). *Tertiarisierung in der deutschen Wirtschaft*. Tübingen: Mohr.

Kocalevent, R. D. (2005). *Entwicklung eines computeradaptiven Tests zur Erfassung von Stresserleben (Stress-CAT)*. Freie Universität Berlin, Berlin. Verfügbar unter http://www.diss.fu-berlin.de/diss/receive/FUDISS_thesis_000000001555

Korschun, D., Bhattacharya, C. B. & Swain, S. D. (2014). Corporate Social Responsibility, Customer Orientation, and the Job Performance of Frontline Employees. *Journal of Marketing*, *78*(3), 20–37. Zugriff am 12.11.2014.

Kruger, J., Wirtz, D. & Miller, D. T. (2005). Counterfactual thinking and the first instinct fallacy. *Journal of Personality and Social Psychology*, *88*(5), 725–735. https://doi.org/10.1037/0022-3514.88.5.725

Kubinger, K. D. (2017). Neue Konzepte und Belege zu den Einsatzmöglichkeiten des AID in der Entwicklungs- und Pädagogischen Psychologie. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *49*(3), 115–126. https://doi.org/10.1026/0049-8637/a000174

Kubinger, K. D. & Spohn, F. (2017). *AID_3_Tailored. Testleiterprogramm zur computergestützten Vorgabe und Auswertung des Adaptiven Intelligenz Diagnostikums 3 (AID 3) von K. D. Kubinger & S. Holocher-Ertl nach dem Prinzip des Tailored Testing*. Göttingen: Hogrefe.

Kubinger, K. D. & Wurst, E. (1985). *Adaptives Intelligenz Diagnostikum (AID)*. Weinheim: Belz.

Kuk, A. Y. C. & Cheng, Y. W. (1997). The monte carlo newton-raphson algorithm. *Journal of Statistical Computation and Simulation*, *59*(3), 233–250. https://doi.org/10.1080/00949657708811858

Ladhari, R. (2008). Alternative measures of service quality. A review. *Managing Service Quality: An International Journal*, *18*(1), 65–86. https://doi.org/10.1108/09604520810842849

Lazarus, A. (2009). Improving Psychiatric Services Through Mystery Shopping. *Psychiatric Services*, *60*(7), 972–973.

Li, C.-H. (2016a). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*(3), 936–949. https://doi.org/10.3758/s13428-015-0619-7

Li, C.-H. (2016b). The performance of ML, DWLS, and ULS estimation with robust corrections in structural equation models with ordinal variables. *Psychological Methods*, *21*(3), 369–387. https://doi.org/10.1037/met0000093

Liao, H. & Chuang, A. (2007). Transforming service employees and climate: A multilevel, multisource examination of transformational leadership in building long-term service relationships. *Journal of Applied Psychology*, *92*(4), 1006–1019. https://doi.org/10.1037/0021-9010.92.4.1006

Liao, H. & Subramony, M. (2008). Employee customer orientation in manufacturing organizations: Joint influences of customer proximity and the senior leadership team. *Journal of Applied Psychology*, *93*(2), 317–328. https://doi.org/10.1037/0021-9010.93.2.317

Lindeberg, J. W. (1922). Eine neue Herleitung des Exponentialgesetztes in der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, *15*, 211–225.

Ling, G., Attali, Y., Finn, B. & Stone, E. A. (2017). Is a Computerized Adaptive Test More Motivating Than a Fixed-Item Test? *Applied Psychological Measurement*, *41*(7), 495–511. https://doi.org/10.1177/0146621617707556

Liu, O. L., Bridgeman, B., Gu, L., Xu, J. & Kong, N. (2015). Investigation of Response Changes in the GRE Revised General Test. *Educational and Psychological Measurement*, *75*(6), 1002–1020. https://doi.org/10.1177/0013164415573988

Locke, E. A. (1969). What is Job Satisfaction? *Organizational Behavior and Human Performance*, *4*(4), 309–336. https://doi.org/10.1016/0030-5073(69)90013-0

MacCallum, R. C. (1995). Model specifications: Procedures, strategies, and related issues. In R. H. Hoyle (Hrsg.), *Structural equation modeling* (S. 16–36). Thousand Oaks, CA: Sage.

Mahlke, J., Schultze, M., Koch, T., Eid, M., Eckert, R. & Brodbeck, F. C. (2016). A Multilevel CFA–MTMM Approach for Multisource Feedback Instruments: Presentation and Application of a New Statistical Model. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(1), 91–110. https://doi.org/10.1080/10705511.2014.990153

Mair, P. & Hatzinger, R. (2007). Extended Rasch Modeling: The eRm Package for the Application of IRT Models in R. *Journal of Statistical Software*, *20*(9), 1–20. https://doi.org/10.18637/jss.v020.i09

Marsh, H. W., Hau, K.-T. & Wen, Z. (2004). In Search of Golden Rules: Comment on Hypothesis-Testing Approaches to Setting Cutoff Values for Fit Indexes and Dangers in Overgeneralizing Hu and Bentler's (1999) Findings. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*(3), 320–341. https://doi.org/10.1207/s15328007sem1103_2

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174.

Maxham, J. G., Netemeyer, R. G. & Lichtenstein, D. R. (2008). The Retail Value Chain: Linking Employee Perceptions to Employee Performance, Customer Evaluations, and Store Performance. *Marketing Science*, *27*(2), 147–167. https://doi.org/10.1287/mksc.1070.0282

McDonald, R. P. (2000). A Basis for Multidimensional Item Response Theory. *Applied Psychological Measurement*, *24*(2), 99–114. https://doi.org/10.1177/01466210022031552

McDonald, R. P. & Ho, M.-H. R. (2002). Principles and Practice in Reporting Structural Equation Analyses. *Psychological Methods*, *7*(1), 64–82. https://doi.org/10.1037//1082-989X.7.1.64

McNeish, D. (2017). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*(3), 412–433. https://doi.org/10.1037/met0000144

Mead, A. D. & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*, *114*(3), 449–458. https://doi.org/10.1037/0033-2909.114.3.449

Michaels, R. E. & Day, R. L. (1985). Measuring Customer Orientation of Salespeople: A Replication with Industrial Buyers. *Journal of Marketing Research*, *22*(4), 443–446. https://doi.org/10.1177/002224378502200409

Minghetti, V. & Celotto, E. (2013). Measuring Quality of Information Services: Combining Mystery Shopping and Customer Satisfaction Research to Assess the Performance of Tourist Offices. *Journal of Travel Research*, *53*(5), 565–580. https://doi.org/10.1177/0047287513506293

Mitra, D. & Golder, P. N. (2006). How Does Objective Quality Affect Perceived Quality? Short-Term Effects, Long-Term Effects, and Asymmetries. *Marketing science*, *25*(3), 230–247. https://doi.org/10.1287/mksc.1050.0175

Moosbrugger, H. & Kelava, A. (Hrsg.). (2012). *Testtheorie und Fragebogenkonstruktion*. Berlin, Heidelberg: Springer Berlin Heidelberg.

Müller, R. & Rupper, P. (Hrsg.). (1994). *Lean-Management in der Praxis: Beiträge zur Gestaltung einer schlanken Produktion*. Zürich: Verlag Industrielle Organisation.

Muraki, E. (1993). Information Functions of the Generalized Partial Credit Model. *Applied Psychological Measurement*, *17*(4), 351–363. https://doi.org/10.1177/014662169301700403

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, *47*(1), 90–100. https://doi.org/10.1016/S0022-2496(02)00028-7

Ogden, S. & Watson, R. (1999). Corporate Performance and Stakeholder Management: Balancing Sharholder and Customer Interests in the U.K. Privatized Water Industry. *Academy of Management Journal*, *42*(5), 526–538. https://doi.org/10.2307/256974

Oliver, R. L. (1980). A Cognitive Model of the Antecedents and Consequences of Satisfaction Decisions. *Journal of Marketing Research*, *17*(4), 460–469. https://doi.org/10.2307/3150499

Oliver, R. L. & Swan, J. E. (1989). Equity and Disconfirmation Perceptions as Influences on Merchant and Product Satisfaction. *Journal of consumer research*, *16*(3), 372–383.

Olson, R., Verley, J., Santos, L. & Salas, C. (2004). What we teach students about the Hawthorne studies: A review of content within a sample of introductory IO and OB textbooks. *The Industrial-Organizational Psychologist*, *41*(3), 23–39.

Ones, D. S., Viswesvaran, C. & Dilchert, S. (2005). Personality at Work: Raising Awareness and Correcting Misconceptions. *Human Performance*, *18*(4), 389–404. https://doi.org/10.1207/s15327043hup1804_5

Ortner, T. M. (2004). On changing the position of items in personality questionnaires: Analysing effects of item sequence using IRT. *Psychology Science*, *46*(4), 466–476.

Ortner, T. M. (2008). Effects of Changed Item Order: A cautionary note to practitioners on jumping to computerized adaptive testing for personality assessment. *International Journal of Selection and Assessment*, *16*(3), 249–257. https://doi.org/10.1111/j.1468-2389.2008.00431.x

Ortner, T. M. & Caspers, J. (2011). Consequences of Test Anxiety on Adaptive Versus Fixed Item Testing. *European Journal of Psychological Assessment*, *27*(3), 157–163. https://doi.org/10.1027/1015-5759/a000062

Ortner, T. M., Weißkopf, E. & Koch, T. (2014). I Will Probably Fail. *European Journal of Psychological Assessment*, *30*(1), 48–56. https://doi.org/10.1027/1015-5759/a000168

Ostini, R. & Nering, M. L. (2006). *Polytomous item response theory models* (Quantitative applications in the social sciences). Thousand Oaks: Sage Publications.

Öztürk, N. B. & Dogan, N. (2015). Investigating Item Exposure Control Methods in Computerized Adaptive Testing. *Educational Sciences: Theory & Practice*, *15*(1), 85–98. https://doi.org/10.12738/estp.2015.1.2593

Pakpahan, E., Hoffmann, R. & Kröger, H. (2017). Statistical methods for causal analysis in life course research: an illustration of a cross-lagged structural equation model, a latent growth model, and an autoregressive latent trajectories model. *International Journal of Social Research Methodology*, *20*(1), 1–19. https://doi.org/10.1080/13645579.2015.1091641

Papanastasiou, E. C. & Reckase, M. D. (2007). A "Rearrangement Procedure" For Scoring Adaptive Tests with Review Options. *International Journal of Testing*, *7*(4), 387–407. https://doi.org/10.1080/15305050701632262

Parasuraman, A. (1994). Alternative scales for measuring service quality: A comparative assessment based on psychometric and diagnostic criteria. *Journal of Retailing*, *70*(3), 201–230. https://doi.org/10.1016/0022-4359(94)90033-7

Parasuraman, A., Berry, L. L. & Zeithaml, V. A. (1991a). Refinement and Reassessment of the SERVQUAL Scale. *Journal of Retailing*, *67*(4), 420–450.

Parasuraman, A., Berry, L. L. & Zeithaml, V. A. (1991b). Understanding, Measuring and Improving Service Quality: Findings from a Multiphase Research Program. In S. W. Brown, E. Gummesson, B. Edvardsson & B. Gustavsson (Hrsg.), *Service Quality: Multidisciplinary and Multinational Perspectives* (S. 253–287). Lexington, Mass: Lexington Books.

Parasuraman, A., Berry, L. L. & Zeithaml, V. A. (1993). More on improving service quality measurement. *Journal of Retailing*, *69*(1), 140–147. https://doi.org/10.1016/S0022-4359(05)80007-7

Parasuraman, A., Zeithaml, V. A. & Berry, L. L. (1985). A conceptual model of service quality and its implications for future research. *Journal of Marketing*, *49*(4), 41–50.

Parasuraman, A., Zeithaml, V. A. & Berry, L. L. (1988). SERVQUAL: A Multiple-Item Scale for Measuring Consumer Perceptions of Service Quality. *Journal of Retailing*, *64*(1), 12–40.

Parasuraman, A., Zeithaml, V. A. & Berry, L. L. (1994). Reassessment of Expectations as a Comparison Standard in Measuring Service Quality: Implications for Further Research. *Journal of Marketing*, *58*(1), 111. https://doi.org/10.2307/1252255

Parasuraman, A., Zeithaml, V. A. & Malhotra, A. (2015). E-S-QUAL. *Journal of Service Research*, *7*(3), 213–233. https://doi.org/10.1177/1094670504271156

Pfeifer, W. (1989). *Etymologisches Wörterbuch des Deutschen. Erarbeitet von einem Autorenkollektiv des Zentralinstituts für Sprachwissenschaft*. Berlin.

Pilkonis, P. A., Choi, S. W., Reise, S. P., Stover, A. M., Riley, W. T. & Cella, D. (2011). Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS®): depression, anxiety, and anger. *Assessment*, *18*(3), 263–283. https://doi.org/10.1177/1073191111411667

Powell, T. & Powell, T. C. (1995). Total Quality Management as Competitive Advantage: A Review and Empirical Study. *Strategic Management Journal*, *16*(1), 15–37. https://doi.org/10.1002/smj.4250160105

Powers, D. E. (2001). Test Anxiety and Test Performance: Comparing Paper-Based and Computer-Adaptive Versions of the Graduate Record Examinations (Gre©) General Test. *Journal of Educational Computing Research*, *24*(3), 249–273. https://doi.org/10.2190/680W-66CR-QRP7-CL1F

R. A language and environment for statistical computing. (2010). Wien: R Foundation for Statistical Computing.

Rafaeli, A., Ziklik, L. & Doucet, L. (2008). The Impact of Call Center Employees' Customer Orientation Behaviors on Service Quality. *Journal of Service Research*, *10*(3), 239–255. https://doi.org/10.1177/1094670507306685

Rapp, R. (1997). *Kundenzufriedenheit durch Servicequalität*. Wiesbaden: Deutscher Universitäts-Verlag.

Rasch, G. (1961). On General Laws and the Meaning of Measurement in Psychology. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability* (Bd. 4, S. 321–333). University of California Press Berkeley, CA.

Revelle, W. (2010). Psych: Procedures for psychological, psychometric, and personality research. *Northwestern University, Evanston, IL, https://personality-project.org/r/psych-manual.pdf*, 1–393.

Revelle, W. & Condon, D. M. (2019). Reliability from α to ω: A tutorial. *Psychological Assessment*, *31*(12), 1395–1411. https://doi.org/10.1037/pas0000754

Rheinberg, F. & Vollmeyer, R. (2003). Flow-Erleben in einem Computerspiel unter experimentell variierten Bedingungen. *Zeitschrift für Psychologie*, *211*(4), 161–170. https://doi.org/10.1026//0044-3409.211.4.161

Roberts, B. & Hogan, R. (Hrsg.). (2001). *Personality psychology in the workplace* (Decade of behavior, 1. Aufl.). Washington, DC: American Psychological Association.

Rod, M. & Ashill, N. J. (2010). The effect of customer orientation on frontline employees job outcomes in a new public management context. *Marketing Intelligence & Planning*, *28*(5), 600–624. https://doi.org/10.1108/02634501011066528

Rodrigues, L. L.R., Barkur, G., Varambally, K.V.M. & Golrooy Motlagh, F. (2011). Comparison of SERVQUAL and SERVPERF metrics. An empirical study. *The TQM Journal*, *23*(6), 629–643. https://doi.org/10.1108/17542731111175248

Rosseel, Y. (2012). lavaan : An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2). https://doi.org/10.18637/jss.v048.i02

Rost, D. & Hoberg, K. (1997). Itempositionsveränderung in Persönlichkeitsfragebogen: Methodischer Kunstfehler oder tolerierbare Praxis? *Diagnostica*, *43*(2), 97–112.

Rozell, E. J., Pettijohn, C. E. & Parker, R. S. (2004). Customer-oriented selling: Exploring the roles of emotional intelligence and organizational commitment. *Psychology and Marketing*, *21*(6), 405–424. https://doi.org/10.1002/mar.20011

Rust, R. T. (2001). The Rise of E-Service. *Journal of Service Research*, *3*(4), 283–284. https://doi.org/10.1177/109467050134001

Rust, R. T. & Oliver, R. L. (1994). Service Quality: Insights and Managerial Implications From the Frontier. In R. T. Rust (ed.), *Service quality. New Directions in Theory and Practice* (S. 1–19). Thousand Oaks, California: Sage Publications.

Rust, R. T. & Zahorik, A. J. (1993). Customer Satisfaction, Customer Retention, and Market Share. *Journal of Retailing*, *69*(2), 193–215. https://doi.org/10.1016/0022-4359(93)90003-2

Ryan, A. M., Schmit, M. J. & Johnson, R. (1996). ATTITUDES AND EFFECTIVENESS: EXAMINING RELATIONS AT AN ORGANIZATIONAL LEVEL. *Personnel Psychology*, *49*(4), 853–882. https://doi.org/10.1111/j.1744-6570.1996.tb02452.x

Salanova, M., Agut, S. & Peiro, J. M. (2005). Linking Organizational Resources and Work Engagement to Employee Performance and Customer Loyalty: The Mediation of Service Climate. *Journal of Applied Psychology*, *90*(6), 1217–1227. https://doi.org/10.1037/0021-9010.90.6.1217

Salvaggio, A. N., Schneider, B., Nishii, L. H., Mayer, D. M., Ramesh, A. & Lyon, J. S. (2007). Manager personality, manager service quality orientation, and service climate: Test of a model. *Journal of Applied Psychology*, *92*(6), 1741–1750. https://doi.org/10.1037/0021-9010.92.6.1741

Sari, H. I. & Raborn, A. (2018). What Information Works Best? A Comparison of Routing Methods. *Applied Psychological Measurement*, *42*(6), 499–515. https://doi.org/10.1177/0146621617752990

Savalei, V. (2014). Understanding Robust Corrections in Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(1), 149–160. https://doi.org/10.1080/10705511.2013.824793

Saxe, R. & Weitz, B. A. (1982). The SOCO scale: a measure of the customer orientation of salespeople. *Journal of Marketing Research*, *19*(3), 343–351.

Schermelleh-Engel, K., Moosbrugger, H. & Müller, H. (2003). Evaluating the Fit of Structural Equation Models: Tests of Significance and Descriptive Goodness-of-Fit Measures. *Methods of Psychological Research Online*, *8*(2), 23–74.

Schneider, B. (1990a). The Climate for Service: An Application of the Climate Construct. In *Organizational climate and culture* (Frontiers of industrial and organizational psychology, 1st ed., S. 383–412). San Francisco: Jossey-Bass.

Schneider, B. (1990b). *Organizational climate and culture* (Frontiers of industrial and organizational psychology, 1. Aufl.). San Francisco: Jossey-Bass.

Schneider, B. & Bowen, D. E. (1985). Employee and customer perceptions of service in banks: Replication and extension. *Journal of Applied Psychology*, *70*(3), 423.

Schneider, B., Bowen, D. E., Ehrhart, M. G. & Holcombe, K. M. (2000). The Climate for Service – Evolution of a Construct. In N. M. Ashkanasy, C. Wilderom & M. F. Peterson (Hrsg.), *Handbook of Organizational Culture and Climate* (S. 21–36). Thousand Oaks, Calif: Sage Publications.

Schneider, B., Ehrhart, M. G. & Macey, W. H. (2013). Organizational Climate and Culture. *Annual Review of Psychology*, *64*(1), 361–388. https://doi.org/10.1146/annurev-psych-113011-143809

Schneider, B., Ehrhart, M. G., Mayer, D. M., Saltz, J. L. & Niles-Jolly, K. (2005). Understanding Organization-Customer Links in Service Settings. *Academy of Management Journal*, *48*(6), 1017–1032. https://doi.org/10.5465/AMJ.2005.19573107

Schneider, B., Holcombe, K. M. & White, S. S. (1997). Lessons learned about service quality: What it is, how to manage it, and how to become a service quality organization. *Consulting Psychology Journal: Practice and Research*, *49*(1), 35–50. https://doi.org/10.1037/1061-4087.49.1.35

Schneider, B., Macey, W. H., Lee, W. C. & Young, S. A. (2009). Organizational Service Climate Drivers of the American Customer Satisfaction Index (ACSI) and Financial and Market Performance. *Journal of Service Research*, *12*(1), 3–14. https://doi.org/10.1177/1094670509336743

Schneider, B., Salvaggio, A. N. & Subirats, M. (2002). Climate Strength: A New Direction for Climate Research. *Journal of Applied Psychology*, *87*(2), 220–229. https://doi.org/10.1037/0021-9010.87.2.220

Schneider, B., White, S. S. & Paul, M. C. (1998). Linking service climate and customer perceptions of service quality: Test of a causal model. *Journal of Applied Psychology*, *83*(2), 150–163. https://doi.org/10.1037/0021-9010.83.2.150

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A. & King, J. (2006). Reporting Structural Equation Modeling and Confirmatory Factor Analysis Results: A Review. *The Journal of Educational Research*, *99*(6), 323–338. https://doi.org/10.3200/JOER.99.6.323-338

Schuh, S. C., Egold, N. W. & van Dick, R. (2012). Towards understanding the role of organizational identification in service settings: A multilevel study spanning leaders, service employees, and customers. *European Journal of Work and Organizational Psychology*, *21*(4), 547–574. https://doi.org/10.1080/1359432X.2011.578391

Schulte, M., Ostroff, C., Shmulyian, S. & Kinicki, A. (2009). Organizational Climate Configurations: Relationships to Collective Attitudes, Customer Satisfaction, and Financial Performance. *Journal of Applied Psychology*, *94*(3), 618–634. https://doi.org/10.1037/a0014365

Sellbom, M. & Tellegen, A. (2019). Factor analysis in psychological assessment research: Common pitfalls and recommendations. *Psychological Assessment*, *31*(12), 1428–1441. https://doi.org/10.1037/pas0000623

Shemwell, D. J. & Yavas, U. (1999). Measuring Service Quality in Hospitals: Scale Development and Managerial Applications. *Journal of Marketing Theory and Practice*, *7*(3), 65–75. https://doi.org/10.1080/10696679.1999.11501841

Smith, M. R., Rasmussen, J. L., Mills, M. J., Wefald, A. J. & Downey, R. G. (2012). Stress and performance: Do service orientation and emotional energy moderate the relationship? *Journal of Occupational Health Psychology*, *17*(1), 116–128. https://doi.org/10.1037/a0026064

Smits, N., Cuijpers, P. & van Straten, A. (2011). Applying computerized adaptive testing to the CES-D scale: a simulation study. *Psychiatry Research*, *188*(1), 147–155. https://doi.org/10.1016/j.psychres.2010.12.001

Snijders, T. A. B. & Roel, J. B. (2012). *Multilevel Analysis: An Introduction To Basic And Advanced Multilevel Modeling* (Methodos series, vol. 12).

Sozialakademie Dortmund. (1999). *Vollbeschäftigung und Tertiarisierung: (Drei-Sektoren-Hypothese): 23. Internationale Tagung der Sozialakademie Dortmund*. Berlin: Duncker & Humblot.

Statistisches Bundesamt. (2020). *Bruttoinlandsprodukt für Deutschland 2019. Begleitmaterial zur Pressekonferenz am 15. Januar 2020 in Berlin*.

Steinbeck, H.-H. (1995). *Das neue Total Quality Management*. Landsberg/Lech: Moderne Industrie.

Strawderman, L. & Koubek, R. (2008). Human Factors and Usability in Service Quality Measurement. *Human Factors and Ergonomics in Manufacturing*, *18*(4), 454–463. https://doi.org/10.1002/hfm.20102

Stürzl, W. (1992). *Lean Production in der Praxis: Spitzenleistungen durch Gruppenarbeit* (Multimind). Paderborn: Junfermann.

Subramony, M. & Pugh, S. D. (2015). Services Management Research: Review, Integration, and Future Directions. *Journal of Management*, *41*(1). https://doi.org/10.1177/0149206314557158

Sundström, A. (2011). Using the Rating Scale Model to Examine the Psychometric Properties of the Self-Efficacy Scale for Driver Competence. *European Journal of Psychological Assessment*, *27*(3), 164–170. https://doi.org/10.1027/1015-5759/a000063

Szymanski, D. M. & Henard, D. H. (2001). Customer Satisfaction: A Meta-Analysis of the Empirical Evidence. *Journal of the Academy of Marketing Science*, *29*(1), 16–35. https://doi.org/10.1177/0092070301291002

Tadepalli, R. (1995). Measuring Customer Orientation of a Salesperson: Modifications of the Soco Scale. *Psychology & Marketing*, *12*(3), 177–187. https://doi.org/10.1002/mar.4220120303

Taylor, S. A. & Baker, T. L. (1994). An assessment of the relationship between service quality and customer satisfaction in the formation of consumers‘ purchase intentions. *Journal of Retailing*, *70*(2), 163–178. https://doi.org/10.1016/0022-4359(94)90013-2

Teas, R. K. (1993). Expectations, Performance Evaluation, and Consumers‘ Perceptions of Quality. *Journal of Marketing*, *57*(4), 18. https://doi.org/10.2307/1252216

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: understanding concepts and applications* (1. Aufl.). Washington, DC: American Psychological Association.

Tonidandel, S., Quiñones, M. A. & Adams, A. A. (2002). Computer-Adaptive Testing: The Impact of Test Characteristics on Perceived Performance and Test Takers' Reactions. *Journal of Applied Psychology*, *87*(2), 320–332. https://doi.org/10.1037//0021-9010.87.2.320

Töpfer, A. & Mehdorn, H. (1995). *Total Quality Management: Anforderungen und Umsetzung im Unternehmen*. Neuwied, Kriftel, Berlin: Hermann Luchterhand Verlag.

TÜV Süd Management Service GmbH. (2017). *Exzellenter Service durch nachhaltige Prozesse. Kriterienkatalog für den Standard ServiceExcellence*. Zugriff am 27.10.2020. Verfügbar unter https://www.tuvsud.com/de-de/-/media/de/management-service/pdf/service-zertifizierungen/tuev-sued-kriterienkatalog-serviceexcellence.pdf

Urban, W. (2013). Perceived quality versus quality of processes: a meta concept of service quality measurement. *The Service Industries Journal*, *33*(2), 200–217. https://doi.org/10.1080/02642069.2011.614337

Van der Linden, Wim J. (2008). Adaptive Models of Psychological Testing. *Zeitschrift für Psychologie / Journal of Psychology*, *216*(1), 1–2. https://doi.org/10.1027/0044-3409.216.1.1

Van der Linden, Wim J. & Ren, H. (2020). A Fast and Simple Algorithm for Bayesian Adaptive Testing. *Journal of Educational and Behavioral Statistics*, *45*(1), 58–85. https://doi.org/10.3102/1076998619858970

Van Doorn, J. (2008). Is There a Halo Effect in Satisfaction Formation in Business-to-Business Services? *Journal of Service Research*, *11*(2), 124–141. https://doi.org/10.1177/1094670508324676

Vaughan, L. & Shiu, E. (2001). ARCHSECRET. A multi-item scale to measure service quality within the voluntary sector. *International Journal of Nonprofit and Voluntary Sector Marketing*, *6*(2), 131–144. https://doi.org/10.1002/nvsm.141

Wainer, H. & Dorans, N. J. (2000). *Computerized Adaptive Testing. A primer* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Walter, O. B., Becker, J., Bjorner, J. B., Fliege, H., Klapp, B. F. & Rose, M. (2007). Development and evaluation of a computer adaptive test for 'Anxiety' (Anxiety-CAT). *Quality of Life Research*, *16*(S1), 143–155. https://doi.org/10.1007/s11136-007-9191-7

Walter, O. B. & Rose, M. (2013). Effect of item order on item calibration and item bank construction for computer adaptive tests. *Psychological Test and Assessment Modeling*, *55*(1), 81–91.

Wang, C. (2013). Mutual Information Item Selection Method in Cognitive Diagnostic Computerized Adaptive Testing With Short Test Length. *Educational and Psychological Measurement*, *73*(6), 1017–1035. https://doi.org/10.1177/0013164413498256

Wang, S., Fellouris, G. & Chang, H.-H. (2019). Statistical Foundations for Computerized Adaptive Testing with Response Revision. *Psychometrika*, *84*(2), 375–394. https://doi.org/10.1007/s11336-019-09662-9

Wang, Y., Lo, H.-P. & Hui, Y. V. (2003). The antecedents of service quality and product quality and their influences on bank reputation. Evidence from the banking industry in China. *Managing Service Quality: An International Journal*, *13*(1), 72–83. https://doi.org/10.1108/09604520310456726

Webber, S. S., Payne, S. C. & Taylor, A. B. (2012). Personality and Trust Fosters Service Quality. *Journal of Business and Psychology*, *27*(2), 193–203. https://doi.org/10.1007/s10869-011-9235-4

Weis, S. & Nuerk, H.-C. (2011). TBS-TK Rezensionen. „FAKT-II. Frankfurter Adaptiver Konzentrationsleistungs-Test". *Psychologische Rundschau*, *62*(2), 139–141. https://doi.org/10.1026/0033-3042/a000063

Weiss, D. J. (1982). Improving Measurement Quality and Efficiency with Adaptive Testing. *Applied Psychological Measurement*, *6*(4), 473–492. https://doi.org/10.1177/014662168200600408

Weiss, D. J. (2011). Better data from better measurements using computerized adaptive testing. *Journal of Methods and Measurement in the Social Sciences*, *2*(1), 1–27. https://doi.org/10.2458/azu_jmmss.v2i1.12351

Wickham, H. (2016). *ggplot2. Elegant graphics for data analysis* (Use R!). Cham: Springer. https://doi.org/10.1007/978-3-319-24277-4

Wildemann, H. (Hrsg.). (1999). *Lean Management: Strategien zur Erreichung wettbewerbsfähiger Unternehmen*. Frankfurt am Main: Frankfurter Allgemeine Zeitung.

Winz, G. (2016). *Qualitätsmanagement für Wirtschaftsingenieure. Qualitätsmethoden, Projektplanung, Kommunikation*. München: Hanser. https://doi.org/10.3139/9783446447684

Wirtz, J. & Bateson, J. E.G. (1999). Consumer Satisfaction with Services: Integrating the Environment Perspective in Services Marketing into the Traditional Disconfirmation Paradigm. *Journal of Business research*, *44*(1), 55–66. https://doi.org/10.1016/S0148-2963(97)00178-1

Wise, S. L., Finney, S. J., Enders, C. K., Freeman, S. A. & Severance, D. D. (1999). Examinee Judgments of Changes in Item Difficulty: Implications for Item Review in Computerized Adaptive Testing. *Applied Measurement in Education*, *12*(2), 185–198. https://doi.org/10.1207/s15324818ame1202_5

Womack, J. P., Jones, D. T. & Roos, D. (1994). *Die zweite Revolution in der Autoindustrie: Konsequenzen aus der weltweiten Studie des Massachusetts Institute of Technology* (Auflage: 1). Frankfurt/Main u. a.: Campus Verlag.

Yagil, D. (2001). Ingratiation and Assertiveness in the Service Provider–Customer Dyad. *Journal of Service Research*, *3*(4), 345–353. https://doi.org/10.1177/109467050134007

Yao, L., Pommerich, M. & Segall, D. O. (2014). Using Multidimensional CAT to Administer a Short, Yet Precise, Screening Test. *Applied Psychological Measurement*, *38*(8), 614–631. https://doi.org/10.1177/0146621614541514

Zablah, A. R., Franke, G. R., Brown, T. J. & Bartholomew, D. E. (2012). How and When Does Customer Orientation Influence Frontline Employee Job Outcomes? A Meta-Analytic Evaluation. *Journal of Marketing*, *76*(3), 21–40. https://doi.org/10.1509/jm.10.0231

Zeithaml, V. A., Parasuraman, A. & Berry, L. L. (1992). *Qualitätsservice: was Ihre Kunden erwarten – was Sie leisten müssen*. Frankfurt/Main, New York: Campus-Verlag.

Zheng, C. & Chang, H.-H. (2016). High-Efficiency Response Distribution-Based Item Selection Algorithms for Short-Length Cognitive Diagnostic Computerized Adaptive Testing. *Applied Psychological Measurement*, *40*(8), 608–624. https://doi.org/10.1177/0146621616665196

Zhongmin, C., Chunyan, L., Yong, H. & Hanwei, C. (2018). *Comparison of Algorithms that Allow Item Review in Computerized Adaptive Testing.* ACT.

Zohar, D. (2000). A group-level model of safety climate: Testing the effect of group climate on microaccidents in manufacturing jobs. *Journal of Applied Psychology*, *85*(4), 587–596. https://doi.org/10.1037//0021-9010.85.4.587

Anhang

Anhang A

Tabelle 9 Item-Kennwerte

Item-Name	Item-Text	*M*	*SD*	$r_{i(t-i)}$	*n*
Frage 1	Mitarbeiter, die sehr guten Service leisten, werden in unserem Unternehmen regelmäßig ausgezeichnet.	2,08	1,07	0,45	891
Frage 2	Das Top-Management berichtet gegenüber allen Mitarbeitern über den Stand der Servicequalität im Unternehmen.	2,76	1,05	0,59	931
Frage 3	Servicequalität und/oder Serviceorientierung ist fest in unserem Unternehmensleitbild verankert.	3,48	0,80	0,55	954
Frage 4	Servicequalität und/oder Serviceorientierung ist fester Bestandteil der Personalentwicklungsmaßnahmen für unsere Führungskräfte.	2,89	1,01	0,63	923
Frage 5	Es existiert ein formalisierter Prozess zur ständigen kundenorientierten Verbesserung unserer Servicequalität.	3,02	1,01	0,58	945
Frage 6	Sehr guter Service wird im Unternehmen als sehr bedeutsam für den wirtschaftlichen Erfolg betrachtet.	3,61	0,69	0,45	959
Frage 7	Unsere Stellenbeschreibungen für Mitarbeiter im direkten Kundenkontakt umfassen Kundenzufriedenheitsziele.	2,54	1,15	0,48	856
Frage 8	Die Dokumentation zu Serviceprozessen ist allen betroffenen Mitarbeitern frei zugänglich.	3,27	1,03	0,44	894
Frage 9	Wir befragen mindestens alle zwei Jahre unsere Kunden zu ihrer Zufriedenheit.	3,39	1,01	0,38	919
Frage 10	Wir ermitteln im Rahmen der Zufriedenheit die Bedeutung von einzelnen Servicemerkmalen.	3,18	1,00	0,50	913
Frage 11	Marktforschung und Bereiche mit direktem Kundenkontakt stimmen sich bei Kundenbefragungen immer ab.	2,70	1,11	0,54	778
Frage 12	Marketing- und Servicebereich stimmen sich zu Leistungsversprechen an Kunden immer ab.	2,98	0,98	0,64	827
Frage 13	Vertriebs- und Servicebereiche treten gegenüber dem Kunden wie "aus einem Guss" auf.	3,09	0,88	0,57	847
Frage 14	Wir fragen unsere Kunden regelmäßig direkt und systematisch nach Ansatzpunkten zur Weiterentwicklung unserer Serviceleistungen.	2,98	0,99	0,59	898

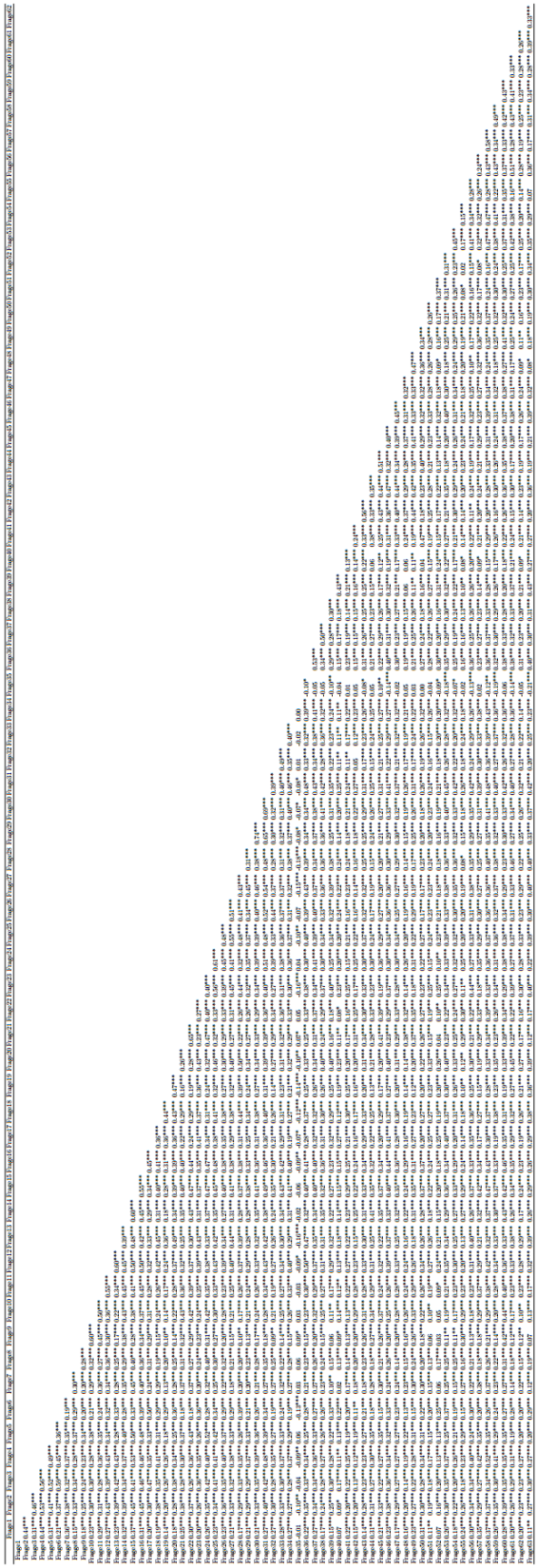
Item-Name	Item-Text	*M*	*SD*	$r_{i(t\text{-}i)}$	*n*
Frage 15	Wir haben einen gelebten Prozess zur Umsetzung von Serviceinnovationen.	2,79	1,01	0,70	881
Frage 16	Es existiert eine kommunizierte Servicestrategie, die Kundenbindung und Kundenzufriedenheit als zentrale Erfolgsfaktoren betrachtet.	3,03	0,98	0,68	868
Frage 17	Kundenzufriedenheit ist ein wesentlicher Bestandteil unseres Unternehmensleitbildes.	3,75	0,58	0,51	894
Frage 18	Wir diskutieren mit unseren Mitarbeitern Prozesse immer aus Kundensicht.	3,09	0,82	0,57	889
Frage 19	Unsere Mitarbeiter wissen, dass gilt: Die Lösung von Kundenanliegen ist wichtiger als die Einhaltung von internen Service-Standards.	3,21	0,81	0,41	875
Frage 20	Unsere Mitarbeiter mit direktem Kundenkontakt versuchen, durch Serviceorientierung für positive Aha-Effekte bei Kunden zu sorgen.	3,35	0,73	0,56	873
Frage 21	Wir nutzen objektiv messbare Kennzahlen und Standards (Wartezeiten, Durchlaufzeiten) für unsere wesentlichen Serviceprozesse zur Messung unserer Servicequalität.	2,81	1,09	0,54	844
Frage 22	Standards bzw. Zielgrößen für unsere Kennzahlen leiten wir immer aus erhobenen Kundenerwartungen ab.	2,61	1,01	0,57	828
Frage 23	Unsere Kunden sind im Durchschnitt mit unserer Servicequalität vollkommen oder sehr zufrieden.	3,44	0,65	0,55	873
Frage 24	Serviceprozesse sind bei uns standardisiert.	3,08	0,92	0,61	854
Frage 25	Wir halten mindestens in 90 Prozent aller Fälle unsere Servicestandards ein.	3,32	0,84	0,63	818
Frage 26	Die Verantwortlichkeit für unsere Serviceprozesse ist immer eindeutig geregelt.	3,32	0,81	0,57	856
Frage 27	Mit unseren vorhandenen personellen und technischen Ressourcen können wir unsere definierten Serviceprozesse immer einhalten.	2,99	0,81	0,58	853
Frage 28	Unsere Servicemitarbeiter versetzen sich so gut wie möglich in die Lage der Kunden und berücksichtigen deren Situation.	3,40	0,71	0,54	847
Frage 29	Unsere vorhandenen IT-Systeme erleichtern unseren Mitarbeitern das Leben unserer Serviceprozesse sehr.	2,98	0,94	0,53	840
Frage 30	Unsere vorhandene Kommunikationstechnik erleichtert unseren Mitarbeitern das Leben unserer Serviceprozesse sehr.	3,16	0,85	0,59	851

Item-Name	Item-Text	*M*	*SD*	$r_{i(t\text{-}i)}$	*n*
Frage 31	Unsere Servicebereiche sind organisatorisch sehr gut verankert bzw. positioniert.	3,14	0,82	0,70	833
Frage 32	Wir kennen die Kosten unserer einzelnen Serviceleistungen genau.	2,81	1,00	0,46	820
Frage 33	Wir erheben oder schätzen die positiven Effekte auf Umsatz und Kundenbindung aus unseren Serviceleistungen.	2,78	0,99	0,54	800
Frage 34	Unsere Mitarbeiter im Servicebereich leiten Cross- und Up-Selling-Chancen direkt an unseren Vertrieb weiter oder realisieren diese selbst.	2,62	1,03	0,54	678
Frage 35	Optimierungen der Wirtschaftlichkeit unserer Serviceleistungen gehen gelegentlich auch auf Kosten der Kundenzufriedenheit.	2,22	1,00	0,08	791
Frage 36	Zu unseren Serviceleistungen kommuniziert unser Marketing ausschließlich, was wir auch im Service einhalten können.	3,02	0,91	0,57	730
Frage 37	Wir kommunizieren explizit unsere Servicestandards und ihre Einhaltung gegenüber unseren Kunden.	3,05	0,90	0,56	772
Frage 38	Unsere Kunden sind sehr gut über unsere Services informiert – auch im Vergleich zum Wettbewerb.	3,16	0,80	0,50	792
Frage 39	Unsere Kunden bekommen immer schnell und direkt Kontakt zu ihren Ansprechpartnern.	3,63	0,62	0,43	830
Frage 40	Unseren Kunden entstehen keine erheblichen zusätzlichen Kosten für die Kontaktaufnahme mit uns.	3,78	0,58	0,27	828
Frage 41	Wir sehen jede Beschwerde/Reklamation tatsächlich als Geschenk des Kunden.	2,84	0,95	0,40	811
Frage 42	Wir haben einen unternehmensweiten Beschwerde-/ Reklamationsprozess im Einsatz.	3,31	0,97	0,39	806
Frage 43	Wir werten Beschwerden/Reklamationen für Verbesserungsmaßnahmen aus.	3,53	0,72	0,54	827
Frage 44	Wir fordern aktiv Kunden über Kommunikationsmaßnahmen auf, ihre Unzufriedenheit als Beschwerden zu artikulieren.	2,74	1,12	0,48	804
Frage 45	Wir messen explizit die Zufriedenheit unserer Kunden mit der Beschwerdebearbeitung.	2,84	1,09	0,51	801
Frage 46	Unser Management erachtet seine regelmäßige Information über aktuelle Beschwerden als sehr wichtig.	3,43	0,82	0,60	818

Item-Name	Item-Text	*M*	*SD*	$r_{i(t-i)}$	*n*
Frage 47	Wir konnten in der Vergangenheit unseren Service auf Basis von Beschwerden an einigen Punkten substanziell weiterentwickeln.	3,18	0,87	0,53	799
Frage 48	Für die Bearbeitung von Beschwerden haben wir Zeitstandards definiert.	2,40	1,25	0,40	802
Frage 49	Unsere Mitarbeiter haben Zugriff auf Leitlinien zum Umgang mit bzw. Verhalten bei Beschwerden.	2,99	1,13	0,49	791
Frage 50	Wir bedanken uns grundsätzlich bei unseren Kunden, die eine Beschwerde an uns richten.	2,63	1,12	0,48	786
Frage 51	Bei Beschwerden erhalten unsere Kunden immer Kontaktdaten von einem persönlichen Ansprechpartner für Rückfragen.	3,38	0,91	0,40	808
Frage 52	Antworten auf Beschwerden werden immer individuell erstellt.	3,64	0,68	0,31	819
Frage 53	Die Qualifikation unserer Mitarbeiter im direkten Kundenkontakt betrachten wir als ganz entscheidende Voraussetzung für einen überdurchschnittlichen und zuverlässigen Service.	3,55	0,71	0,56	812
Frage 54	Unsere Mitarbeiter haben und nutzen definierte Entscheidungsspielräume für Lösungen im Umgang mit Kunden (Empowerment).	3,27	0,82	0,49	794
Frage 55	Wir führen mindestens alle zwei Jahre eine Mitarbeiterzufriedenheitsbefragung durch.	2,61	1,30	0,36	770
Frage 56	Unsere Mitarbeiter sind im Schnitt äußerst oder sehr zufrieden an ihrem Arbeitsplatz.	3,12	0,74	0,51	752
Frage 57	Für die Auswahl unserer Mitarbeiter im direkten Kundenkontakt haben wir bewährte Verfahren zur Eignungsprüfung im Einsatz.	2,42	1,10	0,51	726
Frage 58	Die Weiterentwicklung unserer Mitarbeiter im direkten Kundenkontakt umfasst in hohem Umfang auch Elemente der sozialen Kompetenz.	2,98	0,96	0,65	758
Frage 59	Unsere Mitarbeiter im direkten Kundenkontakt haben zeitlichen Freiraum, um sich untereinander zu unseren gelebten Serviceprozessen auszutauschen.	2,99	0,93	0,55	763
Frage 60	Wir kommunizieren unseren Mitarbeiter im direkten Kundenkontakt Ergebnisse aus unseren Kundenzufriedenheitsbefragungen.	3,27	0,94	0,61	761
Frage 61	Unsere Mitarbeiter im direkten Kundenkontakt sind sehr motiviert, sehr guten Service zu leisten.	3,51	0,64	0,52	788

Anhang B

Tabelle 10 Interkorrelation aller Items der SQ-Skala (größer unter: https://promotion.creaval.de)



Anhang C

Tabelle 11 Standardisierte Item-Parameter aus der Konfirmatorischen Faktorenanalyse

Latente Dimension	Item	Item-Text	β	*SE*	*z*-Wert	*p*-Wert
Service-Kultur	Frage 1	Mitarbeiter, die sehr guten Service leisten werden in unserem Unternehmen regelmäßig ausgezeichnet.	0,59	0,04	14,54	< 0,001
Service-Kultur	Frage 2	Das Top-Management berichtet gegenüber allen Mitarbeitern über den Stand der Servicequalität im Unternehmen.	0,64	0,04	18,31	< 0,001
Service-Kultur	Frage 3	Servicequalität und/oder Serviceorientierung ist fest in unserem Unternehmensleitbild verankert.	0,71	0,04	17,54	< 0,001
Service-Kultur	Frage 4	Servicequalität und/oder Serviceorientierung ist fester Bestandteil der Personalentwicklungsmaßnahmen für unsere Führungskräfte.	0,73	0,03	25,53	< 0,001
Service-Kultur	Frage 5	Es existiert ein formalisierter Prozess zur ständigen kundenorientierten Verbesserung unserer Servicequalität.	0,69	0,04	19,55	< 0,001
Service-Kultur	Frage 6	Sehr guter Service wird im Unternehmen als sehr bedeutsam für den wirtschaftlichen Erfolg betrachtet.	0,61	0,05	12,17	< 0,001
Service-Kultur	Frage 7	Unsere Stellenbeschreibungen für Mitarbeiter im direkten Kundenkontakt umfassen Kundenzufriedenheitsziele.	0,53	0,04	12,73	< 0,001
Service-Kultur	Frage 8	Die Dokumentation zu Serviceprozessen ist allen betroffenen Mitarbeitern frei zugänglich.	0,54	0,05	11,87	< 0,001
Service-Kultur	Frage 9	Wir befragen mindestens alle zwei Jahre unsere Kunden zu ihrer Zufriedenheit.	0,51	0,05	10,08	< 0,001
Service-Kultur	Frage 10	Wir ermitteln im Rahmen der Zufriedenheit die Bedeutung von einzelnen Servicemerkmalen.	0,65	0,04	17,15	< 0,001
Service-Kultur	Frage 11	Marktforschung und Bereiche mit direktem Kundenkontakt stimmen sich bei Kundenbefragungen immer ab.	0,64	0,03	19,34	< 0,001
Service-Kultur	Frage 12	Marketing- und Servicebereich stimmen sich zu Leistungsversprechen an Kunden immer ab.	0,80	0,02	32,57	< 0,001
Service-Kultur	Frage 13	Vertriebs- und Servicebereiche treten gegenüber dem Kunden wie „aus einem Guss“ auf.	0,70	0,03	22,34	< 0,001
Service-Kultur	Frage 14	Wir fragen unsere Kunden regelmäßig direkt und systematisch nach Ansatzpunkten zur Weiterentwicklung unserer Serviceleistungen.	0,69	0,03	21,17	< 0,001
Service-Kultur	Frage 15	Wir haben einen gelebten Prozess zur Umsetzung von Serviceinnovationen.	0,80	0,02	34,41	< 0,001

Latente Dimension	Item	Item-Text	β	*SE*	*z*-Wert	*p*-Wert
Service-Kultur	Frage 16	Es existiert eine kommunizierte Servicestrategie, die Kundenbindung und Kundenzufriedenheit als zentrale Erfolgsfaktoren betrachtet.	0,78	0,03	28,78	< 0,001
Service-Kultur	Frage 17	Kundenzufriedenheit ist ein wesentlicher Bestandteil unseres Unternehmensleitbildes.	0,77	0,04	18,89	< 0,001
Service-Kultur	Frage 18	Wir diskutieren mit unseren Mitarbeitern Prozesse immer aus Kundensicht.	0,68	0,03	21,80	< 0,001
Service-Kultur	Frage 19	Unsere Mitarbeiter wissen, dass gilt: Die Lösung von Kundenanliegen ist wichtiger als die Einhaltung von internen Service-Standards.	0,55	0,04	12,99	< 0,001
Service-Kultur	Frage 20	Unsere Mitarbeiter mit direktem Kundenkontakt versuchen, durch Serviceorientierung für positive Aha-Effekte bei Kunden zu sorgen.	0,69	0,03	21,25	< 0,001
Service-Zuverlässigkeit	Frage 21	Wir nutzen objektiv messbare Kennzahlen und Standards (Wartezeiten, Durchlaufzeiten) für unsere wesentlichen Serviceprozesse zur Messung unserer Servicequalität.	0,64	0,04	17,95	< 0,001
Service-Zuverlässigkeit	Frage 22	Standards bzw. Zielgrößen für unsere Kennzahlen leiten wir immer aus erhobenen Kundenerwartungen ab.	0,66	0,03	18,76	< 0,001
Service-Zuverlässigkeit	Frage 23	Unsere Kunden sind im Durchschnitt mit unserer Servicequalität vollkommen oder sehr zufrieden.	0,74	0,03	22,95	< 0,001
Service-Zuverlässigkeit	Frage 24	Serviceprozesse sind bei uns standardisiert.	0,72	0,03	22,56	< 0,001
Service-Zuverlässigkeit	Frage 25	Wir halten mindestens in 90 Prozent aller Fälle unsere Servicestandards ein.	0,77	0,03	26,63	< 0,001
Service-Zuverlässigkeit	Frage 26	Die Verantwortlichkeit für unsere Serviceprozesse ist immer eindeutig geregelt.	0,70	0,04	19,86	< 0,001
Service-Zuverlässigkeit	Frage 27	Mit unseren vorhandenen personellen und technischen Ressourcen können wir unsere definierten Serviceprozesse immer einhalten.	0,74	0,03	29,39	< 0,001
Service-Zuverlässigkeit	Frage 28	Unsere Servicemitarbeiter versetzen sich so gut wie möglich in die Lage der Kunden und berücksichtigen deren Situation.	0,69	0,04	18,99	< 0,001
Service-Zuverlässigkeit	Frage 29	Unsere vorhandenen IT-Systeme erleichtern unseren Mitarbeitern das Leben unserer Serviceprozesse sehr.	0,72	0,03	24,19	< 0,001
Service-Zuverlässigkeit	Frage 30	Unsere vorhandene Kommunikationstechnik erleichtert unseren Mitarbeitern das Leben unserer Serviceprozesse sehr.	0,80	0,02	33,48	< 0,001
Service-Zuverlässigkeit	Frage 31	Unsere Servicebereiche sind organisatorisch sehr gut verankert bzw. positioniert.	0,84	0,02	41,53	< 0,001
Service-Zuverlässigkeit	Frage 32	Wir kennen die Kosten unserer einzelnen Serviceleistungen genau.	0,52	0,04	12,68	< 0,001

Latente Dimension	Item	Item-Text	β	*SE*	*z*-Wert	*p*-Wert
Service-Zuverlässigkeit	Frage 33	Wir erheben oder schätzen die positiven Effekte auf Umsatz und Kundenbindung aus unseren Serviceleistungen.	0,64	0,03	18,33	< 0,001
Service-Zuverlässigkeit	Frage 34	Unsere Mitarbeiter im Servicebereich leiten Cross- und Up-Selling-Chancen direkt an unseren Vertrieb weiter oder realisieren diese selbst.	0,60	0,04	15,77	< 0,001
Service-Zuverlässigkeit	Frage 36	Zu unseren Serviceleistungen kommuniziert unser Marketing ausschließlich, was wir auch im Service einhalten können.	0,69	0,03	20,55	< 0,001
Service-Zuverlässigkeit	Frage 37	Wir kommunizieren explizit unsere Servicestandards und ihre Einhaltung gegenüber unseren Kunden.	0,70	0,03	20,04	< 0,001
Service-Zuverlässigkeit	Frage 38	Unsere Kunden sind sehr gut über unsere Services informiert – auch im Vergleich zum Wettbewerb.	0,61	0,04	14,73	< 0,001
Service-Zuverlässigkeit	Frage 39	Unsere Kunden bekommen immer schnell und direkt Kontakt zu ihren Ansprechpartnern.	0,57	0,05	12,10	< 0,001
Service-Zuverlässigkeit	Frage 40	Unseren Kunden entstehen keine erheblichen zusätzlichen Kosten für die Kontaktaufnahme mit uns.	0,39	0,07	5,40	< 0,001
Umgang mit Beschwerden	Frage 41	Wir sehen jede Beschwerde/Reklamation tatsächlich als Geschenk des Kunden.	0,61	0,04	14,67	< 0,001
Umgang mit Beschwerden	Frage 42	Wir haben einen unternehmensweiten Beschwerde-/ Reklamationsprozess im Einsatz.	0,63	0,04	14,33	< 0,001
Umgang mit Beschwerden	Frage 43	Wir werten Beschwerden/Reklamationen für Verbesserungsmaßnahmen aus.	0,81	0,04	23,14	< 0,001
Umgang mit Beschwerden	Frage 44	Wir fordern aktiv Kunden über Kommunikationsmaßnahmen auf, ihre Unzufriedenheit als Beschwerden zu artikulieren.	0,69	0,04	17,31	< 0,001
Umgang mit Beschwerden	Frage 45	Wir messen explizit die Zufriedenheit unserer Kunden mit der Beschwerdebearbeitung.	0,68	0,04	17,29	< 0,001
Umgang mit Beschwerden	Frage 46	Unser Management erachtet seine regelmäßige Information über aktuelle Beschwerden als sehr wichtig.	0,81	0,04	22,84	< 0,001
Umgang mit Beschwerden	Frage 47	Wir konnten in der Vergangenheit unseren Service auf Basis von Beschwerden an einigen Punkten substanziell weiterentwickeln.	0,73	0,04	19,18	< 0,001
Umgang mit Beschwerden	Frage 48	Für die Bearbeitung von Beschwerden haben wir Zeitstandards definiert.	0,59	0,04	13,57	< 0,001
Umgang mit Beschwerden	Frage 49	Unsere Mitarbeiter haben Zugriff auf Leitlinien zum Umgang mit bzw. Verhalten bei Beschwerden.	0,65	0,04	16,32	< 0,001

Latente Dimension	Item	Item-Text	β	*SE*	*z*-Wert	*p*-Wert
Umgang mit Beschwerden	Frage 50	Wir bedanken uns grundsätzlich bei unseren Kunden, die eine Beschwerde an uns richten.	0,69	0,04	18,34	< 0,001
Umgang mit Beschwerden	Frage 51	Bei Beschwerden erhalten unsere Kunden immer Kontaktdaten von einem persönlichen Ansprechpartner für Rückfragen.	0,63	0,05	13,02	< 0,001
Umgang mit Beschwerden	Frage 52	Antworten auf Beschwerden werden immer individuell erstellt.	0,50	0,06	8,01	< 0,001
Mitarbeiter-qualifikation	Frage 53	Die Qualifikation unserer Mitarbeiter im direkten Kundenkontakt betrachten wir als ganz entscheidende Voraussetzung für einen überdurchschnittlichen und zuverlässigen Service.	0,75	0,04	19,62	< 0,001
Mitarbeiter-qualifikation	Frage 54	Unsere Mitarbeiter haben und nutzen definierte Entscheidungsspielräume für Lösungen im Umgang mit Kunden (Empowerment).	0,66	0,04	18,21	< 0,001
Mitarbeiter-qualifikation	Frage 55	Wir führen mindestens alle zwei Jahre eine Mitarbeiterzufriedenheitsbefragung durch.	0,44	0,05	8,69	< 0,001
Mitarbeiter-qualifikation	Frage 56	Unsere Mitarbeiter sind im Schnitt äußerst oder sehr zufrieden an ihrem Arbeitsplatz.	0,67	0,04	19,11	< 0,001
Mitarbeiter-qualifikation	Frage 57	Für die Auswahl unserer Mitarbeiter im direkten Kundenkontakt haben wir bewährte Verfahren zur Eignungsprüfung im Einsatz.	0,64	0,04	18,10	< 0,001
Mitarbeiter-qualifikation	Frage 58	Die Weiterentwicklung unserer Mitarbeiter im direkten Kundenkontakt umfasst in hohem Umfang auch Elemente der sozialen Kompetenz.	0,80	0,02	32,71	< 0,001
Mitarbeiter-qualifikation	Frage 59	Unsere Mitarbeiter im direkten Kundenkontakt haben zeitlichen Freiraum, um sich untereinander zu unseren gelebten Serviceprozessen auszutauschen.	0,66	0,04	17,71	< 0,001
Mitarbeiter-qualifikation	Frage 60	Wir kommunizieren unseren Mitarbeiter im direkten Kundenkontakt Ergebnisse aus unseren Kundenzufriedenheitsbefragungen.	0,76	0,03	23,68	< 0,001
Mitarbeiter-qualifikation	Frage 61	Unsere Mitarbeiter im direkten Kundenkontakt sind sehr motiviert, sehr guten Service zu leisten.	0,73	0,04	20,24	< 0,001

Anhang D

Tabelle 12 Schätzung der Schwellenparameter aus der CFA

Item	Schwelle	Schätzer	*SE*	*z*-Wert	*p*-Wert
Frage 1	t1	-0,39	0,07	-5,78	0,00
Frage 1	t2	0,29	0,07	4,45	0,00
Frage 1	t3	1,01	0,08	12,85	0,00
Frage 2	t1	-1,09	0,08	-13,44	0,00
Frage 2	t2	-0,37	0,07	-5,58	0,00
Frage 2	t3	0,45	0,07	6,71	0,00
Frage 3	t1	-1,97	0,14	-14,07	0,00
Frage 3	t2	-1,27	0,09	-14,39	0,00
Frage 3	t3	-0,42	0,07	-6,30	0,00
Frage 4	t1	-1,35	0,09	-14,68	0,00
Frage 4	t2	-0,48	0,07	-7,11	0,00
Frage 4	t3	0,33	0,07	4,96	0,00
Frage 5	t1	-1,42	0,10	-14,86	0,00
Frage 5	t2	-0,68	0,07	-9,63	0,00
Frage 5	t3	0,11	0,07	1,76	0,08
Frage 6	t1	-2,14	0,16	-13,20	0,00
Frage 6	t2	-1,61	0,11	-15,02	0,00
Frage 6	t3	-0,64	0,07	-9,14	0,00
Frage 7	t1	-0,67	0,07	-9,53	0,00
Frage 7	t2	-0,14	0,07	-2,17	0,03
Frage 7	t3	0,65	0,07	9,24	0,00
Frage 8	t1	-1,44	0,10	-14,90	0,00
Frage 8	t2	-0,90	0,08	-11,95	0,00
Frage 8	t3	-0,23	0,07	-3,52	0,00
Frage 9	t1	-1,37	0,09	-14,73	0,00
Frage 9	t2	-0,91	0,08	-12,04	0,00
Frage 9	t3	-0,46	0,07	-6,81	0,00
Frage 10	t1	-1,50	0,10	-14,98	0,00
Frage 10	t2	-0,83	0,07	-11,19	0,00
Frage 10	t3	-0,03	0,07	-0,52	0,60
Frage 11	t1	-1,02	0,08	-12,93	0,00
Frage 11	t2	-0,31	0,07	-4,65	0,00
Frage 11	t3	0,45	0,07	6,71	0,00
Frage 12	t1	-1,37	0,09	-14,73	0,00
Frage 12	t2	-0,51	0,07	-7,52	0,00

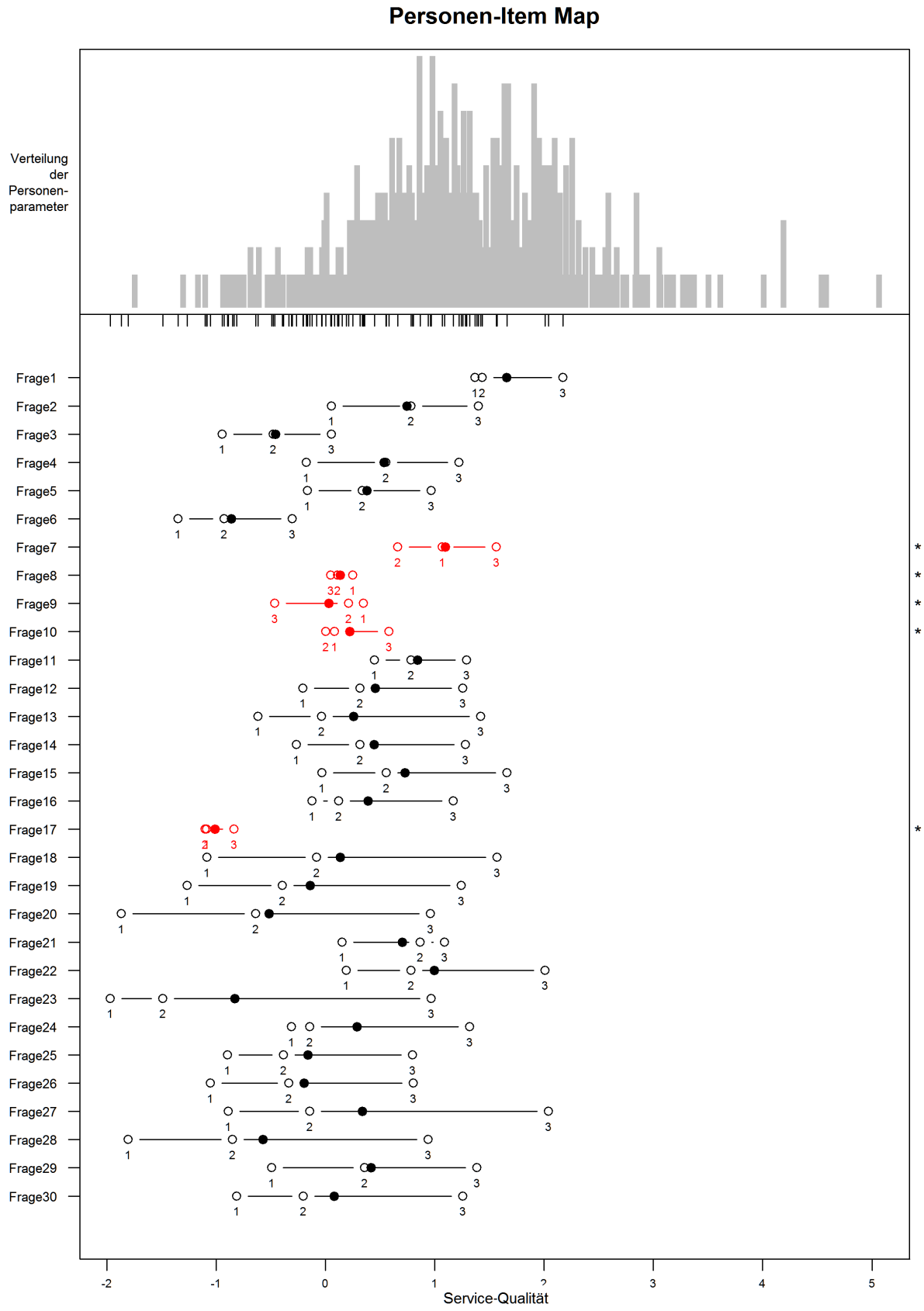
Item	Schwelle	Schätzer	*SE*	*z*-Wert	*p*-Wert
Frage 12	t3	0,36	0,07	5,37	0,00
Frage 13	t1	-1,52	0,10	-15,00	0,00
Frage 13	t2	-0,61	0,07	-8,73	0,00
Frage 13	t3	0,44	0,07	6,50	0,00
Frage 14	t1	-1,37	0,09	-14,73	0,00
Frage 14	t2	-0,54	0,07	-7,93	0,00
Frage 14	t3	0,29	0,07	4,45	0,00
Frage 15	t1	-1,20	0,09	-14,05	0,00
Frage 15	t2	-0,34	0,07	-5,17	0,00
Frage 15	t3	0,56	0,07	8,13	0,00
Frage 16	t1	-1,40	0,09	-14,82	0,00
Frage 16	t2	-0,67	0,07	-9,43	0,00
Frage 16	t3	0,25	0,07	3,72	0,00
Frage 17	t1	-2,21	0,17	-12,75	0,00
Frage 17	t2	-1,69	0,11	-14,94	0,00
Frage 17	t3	-0,85	0,07	-11,38	0,00
Frage 18	t1	-1,69	0,11	-14,94	0,00
Frage 18	t2	-0,76	0,07	-10,52	0,00
Frage 18	t3	0,39	0,07	5,78	0,00
Frage 19	t1	-2,02	0,15	-13,83	0,00
Frage 19	t2	-0,97	0,08	-12,49	0,00
Frage 19	t3	0,29	0,07	4,34	0,00
Frage 20	t1	-2,14	0,16	-13,20	0,00
Frage 20	t2	-1,09	0,08	-13,44	0,00
Frage 20	t3	0,09	0,07	1,45	0,15
Frage 21	t1	-1,17	0,08	-13,90	0,00
Frage 21	t2	-0,31	0,07	-4,65	0,00
Frage 21	t3	0,42	0,07	6,30	0,00
Frage 22	t1	-1,09	0,08	-13,44	0,00
Frage 22	t2	-0,18	0,07	-2,79	0,01
Frage 22	t3	0,82	0,07	11,10	0,00
Frage 23	t1	-2,41	0,21	-11,44	0,00
Frage 23	t2	-1,40	0,09	-14,82	0,00
Frage 23	t3	0,04	0,07	0,62	0,53
Frage 24	t1	-1,61	0,11	-15,02	0,00
Frage 24	t2	-0,82	0,07	-11,10	0,00
Frage 24	t3	0,30	0,07	4,55	0,00
Frage 25	t1	-1,81	0,12	-14,67	0,00

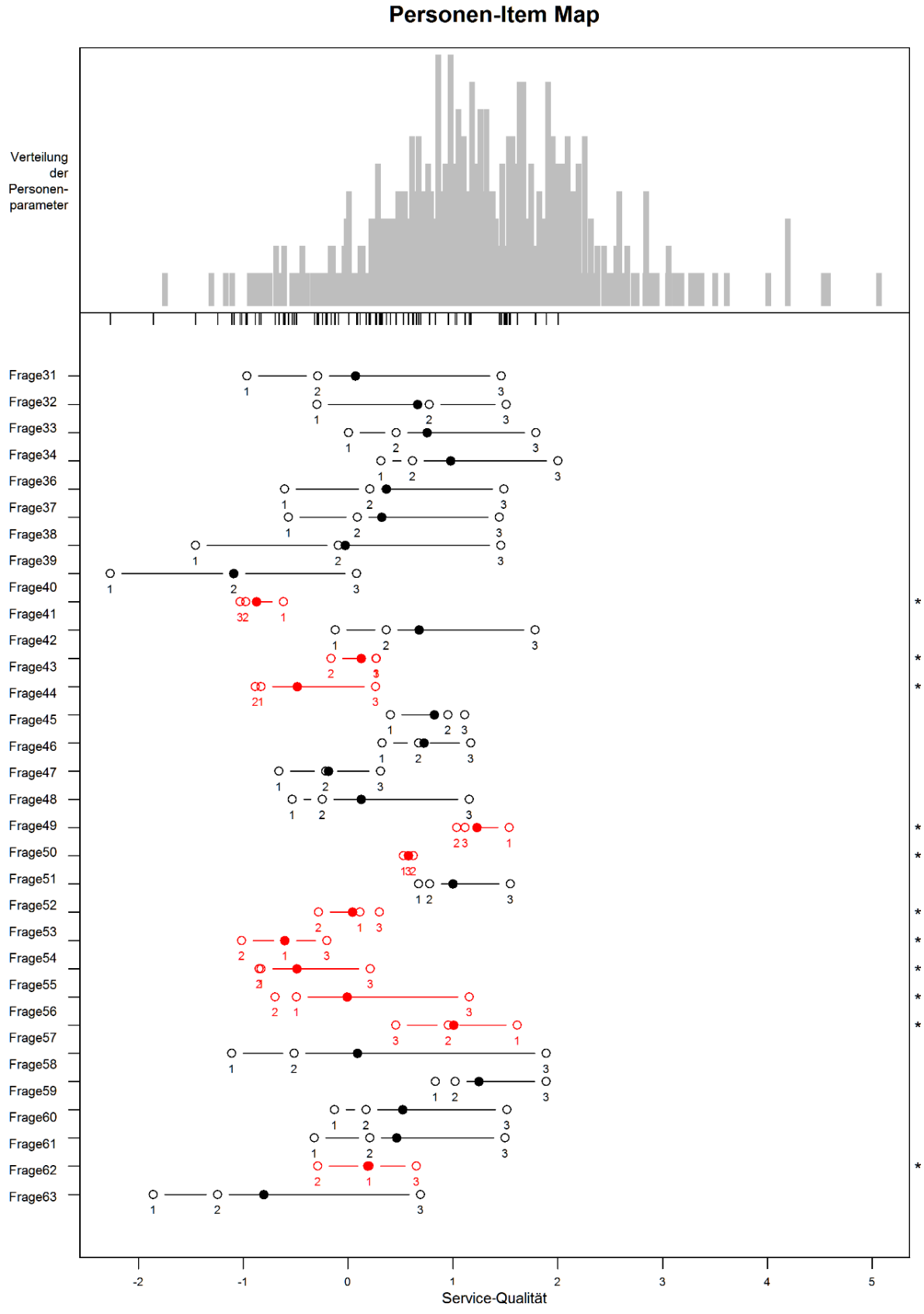
Item	Schwelle	Schätzer	*SE*	*z*-Wert	*p*-Wert
Frage 25	t2	-1,05	0,08	-13,10	0,00
Frage 25	t3	0,02	0,07	0,31	0,76
Frage 26	t1	-1,85	0,13	-14,56	0,00
Frage 26	t2	-1,02	0,08	-12,93	0,00
Frage 26	t3	0,01	0,07	0,21	0,84
Frage 27	t1	-1,61	0,11	-15,02	0,00
Frage 27	t2	-0,64	0,07	-9,14	0,00
Frage 27	t3	0,68	0,07	9,63	0,00
Frage 28	t1	-2,21	0,17	-12,75	0,00
Frage 28	t2	-1,21	0,09	-14,12	0,00
Frage 28	t3	0,00	0,07	0,00	1,00
Frage 29	t1	-1,44	0,10	-14,90	0,00
Frage 29	t2	-0,57	0,07	-8,23	0,00
Frage 29	t3	0,40	0,07	5,99	0,00
Frage 30	t1	-1,81	0,12	-14,67	0,00
Frage 30	t2	-0,89	0,08	-11,85	0,00
Frage 30	t3	0,25	0,07	3,83	0,00
Frage 31	t1	-1,85	0,13	-14,56	0,00
Frage 31	t2	-0,86	0,07	-11,57	0,00
Frage 31	t3	0,33	0,07	4,96	0,00
Frage 32	t1	-1,27	0,09	-14,39	0,00
Frage 32	t2	-0,32	0,07	-4,86	0,00
Frage 32	t3	0,49	0,07	7,22	0,00
Frage 33	t1	-1,25	0,09	-14,33	0,00
Frage 33	t2	-0,45	0,07	-6,60	0,00
Frage 33	t3	0,67	0,07	9,53	0,00
Frage 34	t1	-1,06	0,08	-13,19	0,00
Frage 34	t2	-0,25	0,07	-3,72	0,00
Frage 34	t3	0,83	0,07	11,19	0,00
Frage 36	t1	-1,42	0,10	-14,86	0,00
Frage 36	t2	-0,48	0,07	-7,01	0,00
Frage 36	t3	0,56	0,07	8,13	0,00
Frage 37	t1	-1,61	0,11	-15,02	0,00
Frage 37	t2	-0,62	0,07	-8,84	0,00
Frage 37	t3	0,48	0,07	7,11	0,00
Frage 38	t1	-1,93	0,14	-14,26	0,00
Frage 38	t2	-0,85	0,07	-11,38	0,00
Frage 38	t3	0,41	0,07	6,09	0,00

Item	Schwelle	Schätzer	*SE*	*z*-Wert	*p*-Wert
Frage 39	t1	-2,41	0,21	-11,44	0,00
Frage 39	t2	-1,44	0,10	-14,90	0,00
Frage 39	t3	-0,37	0,07	-5,58	0,00
Frage 40	t1	-2,08	0,15	-13,55	0,00
Frage 40	t2	-1,61	0,11	-15,02	0,00
Frage 40	t3	-0,89	0,08	-11,85	0,00
Frage 41	t1	-1,23	0,09	-14,19	0,00
Frage 41	t2	-0,45	0,07	-6,71	0,00
Frage 41	t3	0,59	0,07	8,53	0,00
Frage 42	t1	-1,52	0,10	-15,00	0,00
Frage 42	t2	-0,97	0,08	-12,49	0,00
Frage 42	t3	-0,18	0,07	-2,69	0,01
Frage 43	t1	-1,97	0,14	-14,07	0,00
Frage 43	t2	-1,32	0,09	-14,57	0,00
Frage 43	t3	-0,30	0,07	-4,55	0,00
Frage 44	t1	-1,00	0,08	-12,76	0,00
Frage 44	t2	-0,30	0,07	-4,55	0,00
Frage 44	t3	0,41	0,07	6,09	0,00
Frage 45	t1	-1,14	0,08	-13,75	0,00
Frage 45	t2	-0,44	0,07	-6,50	0,00
Frage 45	t3	0,37	0,07	5,58	0,00
Frage 46	t1	-1,66	0,11	-14,98	0,00
Frage 46	t2	-0,98	0,08	-12,58	0,00
Frage 46	t3	-0,14	0,07	-2,17	0,03
Frage 47	t1	-1,56	0,10	-15,02	0,00
Frage 47	t2	-0,93	0,08	-12,13	0,00
Frage 47	t3	0,11	0,07	1,76	0,08
Frage 48	t1	-0,51	0,07	-7,52	0,00
Frage 48	t2	0,01	0,07	0,10	0,92
Frage 48	t3	0,51	0,07	7,52	0,00
Frage 49	t1	-1,12	0,08	-13,60	0,00
Frage 49	t2	-0,53	0,07	-7,72	0,00
Frage 49	t3	0,14	0,07	2,17	0,03
Frage 50	t1	-0,88	0,08	-11,76	0,00
Frage 50	t2	-0,27	0,07	-4,03	0,00
Frage 50	t3	0,57	0,07	8,23	0,00
Frage 51	t1	-1,46	0,10	-14,93	0,00
Frage 51	t2	-0,98	0,08	-12,58	0,00

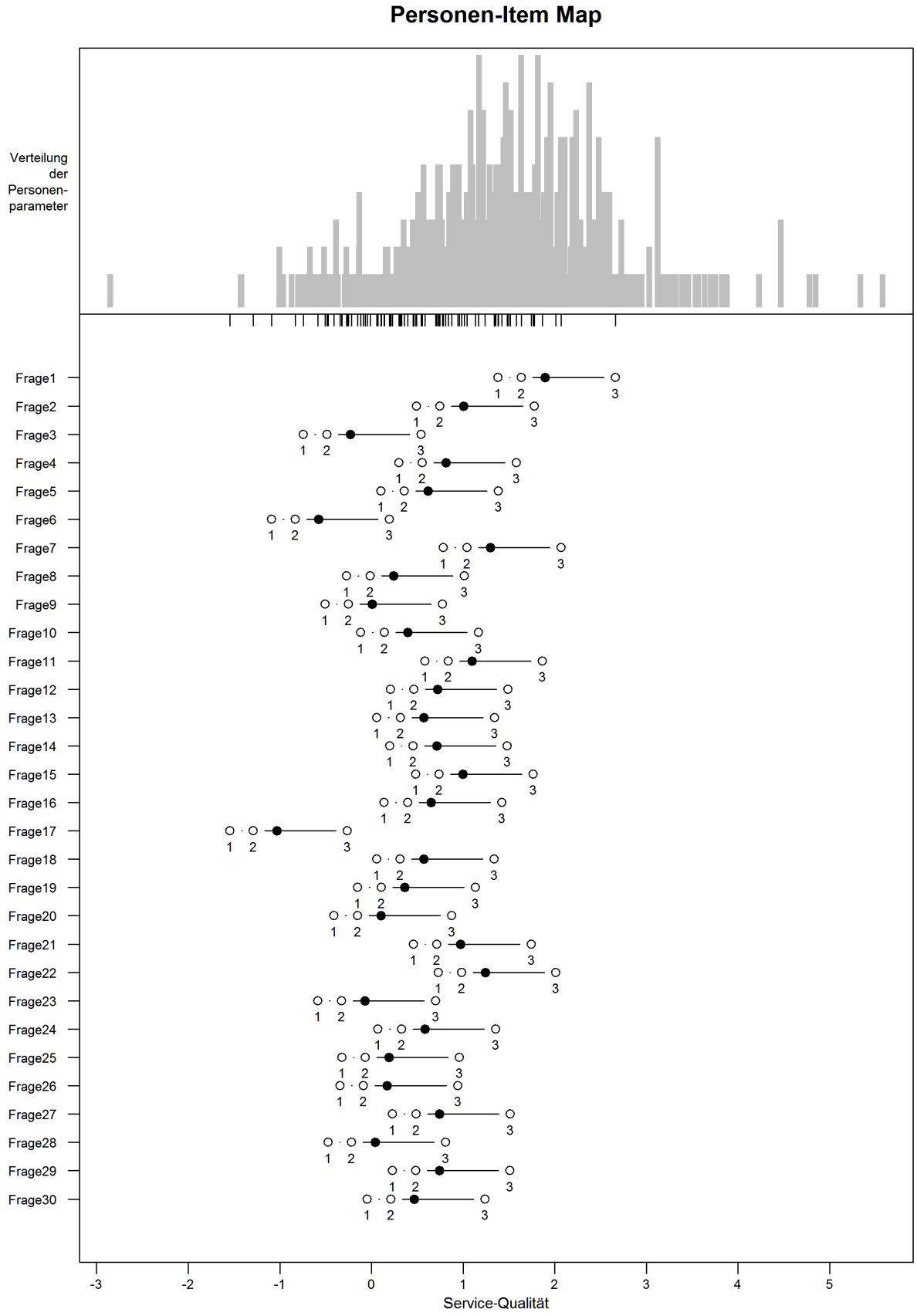
Item	Schwelle	Schätzer	*SE*	*z*-Wert	*p*-Wert
Frage 51	t3	-0,21	0,07	-3,21	0,00
Frage 52	t1	-1,89	0,13	-14,43	0,00
Frage 52	t2	-1,33	0,09	-14,63	0,00
Frage 52	t3	-0,53	0,07	-7,72	0,00
Frage 53	t1	-1,85	0,13	-14,56	0,00
Frage 53	t2	-1,20	0,09	-14,05	0,00
Frage 53	t3	-0,35	0,07	-5,27	0,00
Frage 54	t1	-1,75	0,12	-14,84	0,00
Frage 54	t2	-1,06	0,08	-13,19	0,00
Frage 54	t3	0,12	0,07	1,86	0,06
Frage 55	t1	-0,60	0,07	-8,63	0,00
Frage 55	t2	-0,20	0,07	-3,00	0,00
Frage 55	t3	0,22	0,07	3,41	0,00
Frage 56	t1	-1,72	0,12	-14,90	0,00
Frage 56	t2	-0,86	0,07	-11,57	0,00
Frage 56	t3	0,54	0,07	7,83	0,00
Frage 57	t1	-0,78	0,07	-10,71	0,00
Frage 57	t2	-0,05	0,07	-0,83	0,41
Frage 57	t3	0,84	0,07	11,29	0,00
Frage 58	t1	-1,35	0,09	-14,68	0,00
Frage 58	t2	-0,58	0,07	-8,33	0,00
Frage 58	t3	0,45	0,07	6,71	0,00
Frage 59	t1	-1,48	0,10	-14,96	0,00
Frage 59	t2	-0,55	0,07	-8,03	0,00
Frage 59	t3	0,45	0,07	6,60	0,00
Frage 60	t1	-1,48	0,10	-14,96	0,00
Frage 60	t2	-0,98	0,08	-12,58	0,00
Frage 60	t3	-0,04	0,07	-0,62	0,53
Frage 61	t1	-2,30	0,19	-12,19	0,00
Frage 61	t2	-1,40	0,09	-14,82	0,00
Frage 61	t3	-0,09	0,07	-1,45	0,15

Anhang E

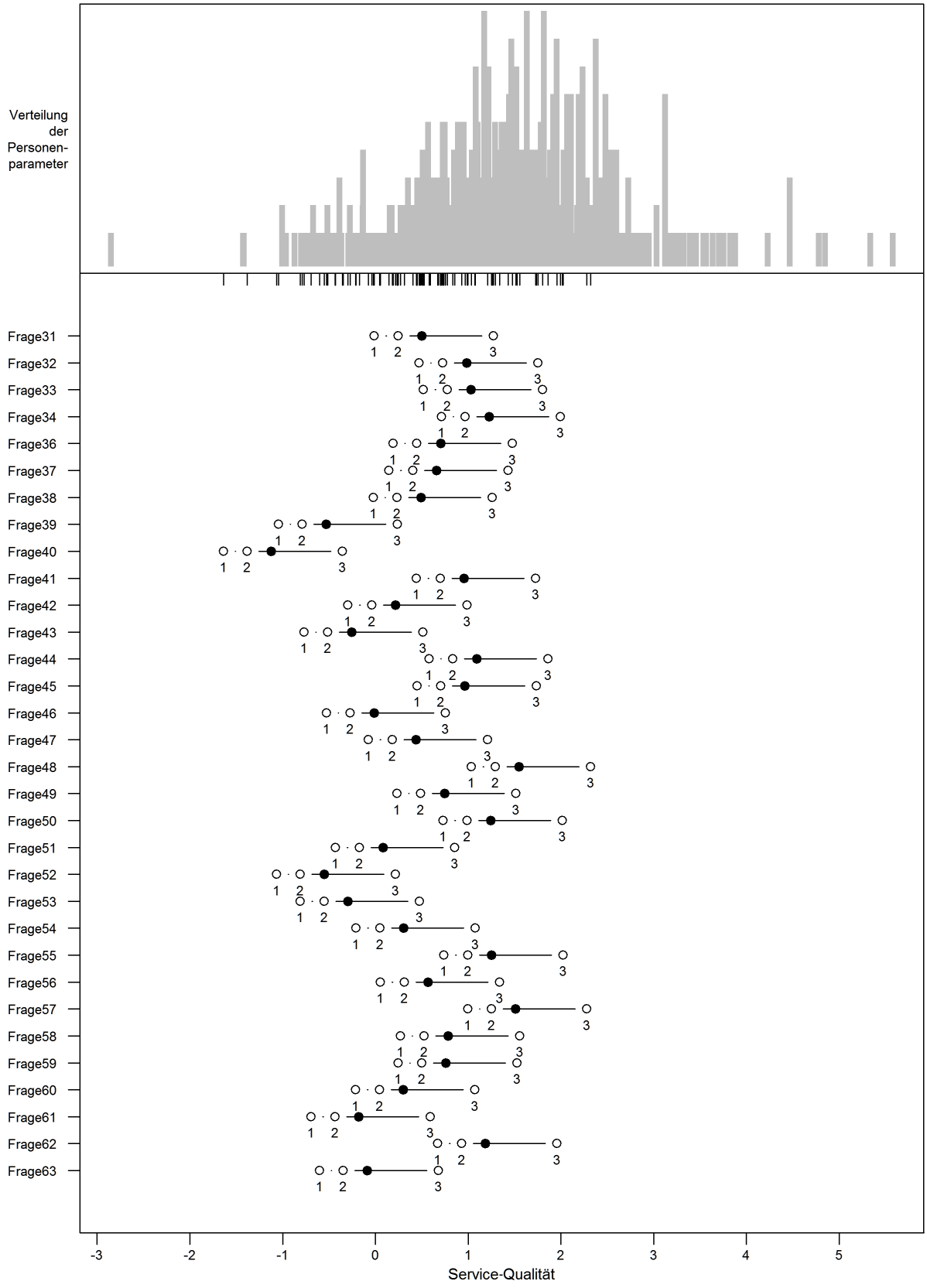

Personen-Item Map
Verteilung der Personen-parameter
Frage1
Frage2
Frage3
Frage4
Frage5
Frage6
Frage7
Frage8
Frage9
Frage10
Frage11
Frage12
Frage13
Frage14
Frage15
Frage16
Frage17
Frage18
Frage19
Frage20
Frage21
Frage22
Frage23
Frage24
Frage25
Frage26
Frage27
Frage28
Frage29
Frage30
-2
-1
0
1
2
3
4
5
Service-Qualität


Personen-Item Map
Verteilung der Personen-parameter
Frage31
Frage32
Frage33
Frage34
Frage36
Frage37
Frage38
Frage39
Frage40
Frage41
Frage42
Frage43
Frage44
Frage45
Frage46
Frage47
Frage48
Frage49
Frage50
Frage51
Frage52
Frage53
Frage54
Frage55
Frage56
Frage57
Frage58
Frage59
Frage60
Frage61
Frage62
Frage63
-2
-1
0
1
2
3
4
5
Service-Qualität

Anhang F


Personen-Item Map
Verteilung der Personen-parameter
Frage1
Frage2
Frage3
Frage4
Frage5
Frage6
Frage7
Frage8
Frage9
Frage10
Frage11
Frage12
Frage13
Frage14
Frage15
Frage16
Frage17
Frage18
Frage19
Frage20
Frage21
Frage22
Frage23
Frage24
Frage25
Frage26
Frage27
Frage28
Frage29
Frage30
-3
-2
-1
0
1
2
3
4
5
Service-Qualität

Personen-Item Map



Eidesstattliche Erklärung

Hiermit versichere ich, dass ich diese Dissertation selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen benutzt habe.

Des Weiteren bestätige ich, dass ich die Dissertation oder Teile hiervon noch nicht als Prüfungsarbeit für eine staatliche oder wissenschaftliche Prüfung sowie noch an keiner anderen Hochschule eingereicht habe.

Landau 03.11.2020

Andreas Pfeiffer

Lebenslauf

Persönliche Daten

Name	Andreas Pfeiffer
geboren am	22.01.1982
Geburtsort	Lörrach
Telefon	0176 66 00 22 64
E-Mail	Andreas.Pfeiffer@gmx.de

Schulische Ausbildung

1993 – 1998	Realschule in Weil am Rhein
1998 – 2001	Technisches Gymnasium in Lörrrach
09/2002 – 03/2010	Diplom Psychologie an der Universität Koblenz – Landau

Beruflicher Werdegang

09/2003 – 03/2010	Wissenschaftliche Hilfskraft: Tutor für Methodenlehre und Versuchsplanung
06/2010 – 12/2010	Wissenschaftlicher Mitarbeiter am Institut für Sonderpädagogik der Universität Koblenz-Landau
01/2011 – 12/2020	Wissenschaftlicher Mitarbeiter am Methodenzentrum der Universität Koblenz-Landau