

Vorhersagekraft von Noten in Verbraucherschutztests auf die Kundenzufriedenheit – eine empirische Untersuchung am Beispiel von Pkw in Deutschland¹

Jan Hendrik Schreier

ADAC e. V., Otto-Lilienthal-Str. 2, 86899 Landsberg am Lech,
jan.schreier@tzll.adac.de.

Prof. Dr. Niels Biethahn

Institut für Automobil Forschung (im RIF e.V.) und Professor für Unternehmenssteuerung und Projekt Manager für Automotive Management, BITS Business and Information Technology School GmbH, Reiterweg 26b, 58535 Iserlohn, Niels.Biethahn@bits-iserlohn.de.

1	Einführung und Problemstellung	171
2	Bestehende Forschung zur Erfolgskontrolle von Warentests und Gewichtung von Produktattributen.....	173
3	Vorstellung der Kundenzufriedenheitsstudie und des Autotests	177
4	Abgleich zwischen Autotest und Kundenzufriedenheitsstudie	182
5	Fazit und Ausblick	189
6	Literaturverzeichnis	190

Abstract

Verbraucherschutztests sollen Konsumenten bei der Kaufentscheidung unterstützen, indem sie auf dem Markt verfügbare Produkte bewerten und miteinander vergleichen. In dieser Veröffentlichung werden die Ergebnisse einer empirischen Untersuchung zur Vorhersagekraft von Bewertungen aus Verbraucherschutztests für die Kundenzufriedenheit am Beispiel von Personenkraftwagen vorgestellt. Analysiert wird zunächst der Zusammenhang zwischen Gesamtnote und Kundenzufriedenheit und dann die Beziehungen von Teilnoten und Zufriedenheit mit den Teilaspekten. Der betrachtete Verbraucherschutztest ist der ADAC-Autotest. In der vorliegenden Veröffentlichung wird zunächst ein Verfahren zum Abgleich von Verbraucherschutztests und Zufriedenheitsbefragungen hergeleitet und angewendet. Hauptergebnis des Abgleichs ist, dass der Zusammenhang zwischen Autotest und Kundenzufriedenheit verhältnismäßig hoch ist. Es besteht jedoch auch noch weiterhin Potential für Verbesserungen. Der Zusammenhang von Teilnoten und Teilzufriedenheiten ist in einigen Fällen sehr niedrig und sollte weiter untersucht werden.

JEL Classification: M31 and O39

Keywords: Customer Satisfaction, Automotive, Consumer Protection, Expert Tests

¹ Die Ergebnisse, Interpretationen und Schlussfolgerungen in diesem Aufsatz sind ausschließlich die Meinung der Autoren. Sie repräsentieren nicht die Meinung des ADAC e. V.. Dank geht an Frank Drewes und die Harris Interactive AG für die Unterstützung bei der statistischen Analyse.

1 Einführung und Problemstellung

1.1 *Fehlende Erfolgskontrollen für Warentests*

Um Verbraucher bei Kaufentscheidungen zu unterstützen, sollen Verbraucherschutztests Konsumenten vor dem Kauf über Qualität und wesentliche Eigenschaften der verfügbaren Produkte informieren (Stiftung Warentest, 2011b, § 2(1)). Dies setzt unter anderem voraus, dass Verbraucher von der Existenz der Tests wissen und die Ergebnisse auch in einer Form vorliegen, die den Anforderungen der Verbraucher entspricht.

Die großen Verbraucherschutzorganisationen in Deutschland verfügen über einen hohen Bekanntheitsgrad (Stiftung Warentest, Forsa, 2007). Dieser reichte jedoch in der Vergangenheit nicht dazu, dass aus Sicht der Organisationen alle Testergebnisse weit genug verbreitet wurden. Mittels der Veröffentlichung der Ergebnisse in kompakter Form konnte z. B. die Stiftung Warentest eine höhere Verbreitung durch andere Medien und somit eine wesentlich höhere Gesamtreichweite erzielen. Ein wesentlicher Teil dieser Zusammenfassung ist die Verdichtung der Ergebnisse auf eine Gesamtnote je Produkt (Laak, 2004: S. 32–34).

Gesamtnoten erlauben Verbrauchern eine schnelle Informationsaufnahme und kommen so einer wesentlichen Anforderung vieler Verbraucher nach (Lösenbeck, 2003: S. 100; Silberer, 1985: S. 10). Gesamtnoten haben jedoch den Nachteil, dass verschiedene Produkteigenschaften gewichtet werden müssen, damit eine Gesamtnote berechnet werden kann. Es ist anzunehmen, dass nicht alle Konsumenten die gleichen Erwartungen an ein Produkt haben. Die Gesamtnote ist daher zumeist ein Kompromiss mit dem Ziel, die Durchschnittsmeinung der Mehrheit der Verbraucher widerzuspiegeln. Bisher ist wenig erforscht, inwieweit die Gesamtnote dieses Ziel erfüllt. Wie groß der Anteil der Verbraucher ist, für den die Gewichtung nicht dem eigenen Urteil entspricht, ist ebenfalls unbekannt (siehe Kapitel 2).

Nicht nur die Gewichtungen, auch die Bewertungsschemata von Haupt-, Unter- und Einzelkriterien² wurden bisher keiner Überprüfung durch empirische Daten unterzogen. Zwar gibt es für viele Einzelkriterien nahezu eindeutige Anforderungen³, für manche Kriterien ist dies jedoch nicht der Fall⁴. Es ist zu prüfen, ob insbesondere in letzterem Fall Bewertungskriterien aus Sicht der Verbraucher in aktuellen Warentests korrekt definiert sind.

Das Unwissen darüber, ob die Gewichte und Bewertungen der Einzelkriterien den Anforderungen der Mehrheit der Konsumenten entsprechen, ist aus Sicht der Warentestorganisationen kritisch. Fehlerhafte Bewertungen und Gewichtungen können dazu führen, dass Verbraucher sich für ein Produkt entscheiden, welches ihre Anforderungen nur unzureichend erfüllt. Indirekt kann ein so entstandener Fehlkauf dazu führen,

² Unter einem Einzelkriterium wird eine konkret bewertbare Eigenschaft eines Produkts verstanden. Aus Gründen der Übersicht können mehrere Einzel- zu Unterkriterien und diese wiederum zu Hauptkriterien zusammengefasst werden. Aus letzteren wird dann die Gesamtnote ermittelt.

³ Ein Fahrzeug sollte z. B. möglichst wenig Kraftstoff verbrauchen.

⁴ Die optimale Anordnung der Bedienelemente in einem Fahrzeug ist z. B. schwer zu definieren bzw. bisher nicht eindeutig definiert.

dass Verbraucher sich weniger auf Warentests verlassen und der Stellenwert dieser im Besonderen und des Verbraucherschutzes im Allgemeinen sinkt.

Da die meisten Produkte im Laufe der Zeit von den Herstellern weiterentwickelt werden, müssen auch Produkttests mit der Zeit angepasst werden. Würde dies nicht geschehen, hätten im Laufe der Zeit alle Produkte die Bestnote oder wichtige neue Funktionen blieben im Test unberücksichtigt. Zusätzlich ändert sich in der Regel das Anforderungsprofil an Produkte: Funktionen, die einstmals für Begeisterung sorgten, werden nach einiger Zeit zum Standard. Dies führt dazu, dass deren Fehlen vom Kunden bemängelt wird, ihr Vorhandensein jedoch keine Begeisterung mehr auslöst (Brandt, 1988: S. 40). Derartige Entwicklungen, rechtliche Vorgaben und auch Trends müssen in Verbraucherschutztests berücksichtigt werden, damit die Tests Verbrauchern einen Nutzen liefern und dadurch relevant bleiben.

Bisher existieren allerdings nur wenige wissenschaftliche Untersuchungen, die den Erfolg von Verbraucherschutztests hinsichtlich ihres Ziels „Verbraucherinformation“ messen (siehe hierzu das Kapitel 0).

1.2 Ziel und Gang der Untersuchung

Das Ziel dieser Untersuchung ist es, exemplarisch am Beispiel des ADAC-Autotests⁵ und mittels der Daten der ADAC-Kundenzufriedenheitsstudie (ADAC, 2012b) einen Abgleich zwischen den Noten des Autotests und der Verbrauchierzufriedenheit durchzuführen. Dadurch wird für Gesamtnoten und Hauptkriterien Aufschluss darüber gewonnen, wie stark die Ergebnisse des Autotests und die gemessene Kundenzufriedenheit voneinander abweichen. Dies liefert Hinweise, ob und wenn ja, wie groß der Handlungsdruck bezüglich der Weiterentwicklung von Verfahren zur Entwicklung von Warentests ist.

Nach der Einführung in diesem Kapitel werden im zweiten Kapitel bestehende Untersuchungen zur Erfolgsmessung von Warentests vorgestellt. Der überprüfte Autotest und die verwendete Kundenzufriedenheitsstudie werden im dritten Kapitel beschrieben. Das vierte Kapitel befasst sich mit der Herleitung und der Durchführung des Abgleichs zwischen Warentest und Zufriedenheitsstudie. Im fünften Kapitel werden die Ergebnisse kritisch diskutiert und weiterer Forschungsbedarf aufgezeigt. Der Aufbau der Untersuchung ist in Abbildung 12 dargestellt. Die Pfeile zeigen an, welche Kapitel logisch direkt aufeinander aufbauen.

⁵ Der Allgemeine Deutsche Automobil-Club e. V. (ADAC) ist Deutschlands größter Automobilclub mit über 18 Millionen Mitgliedern (ADAC, 2012a).

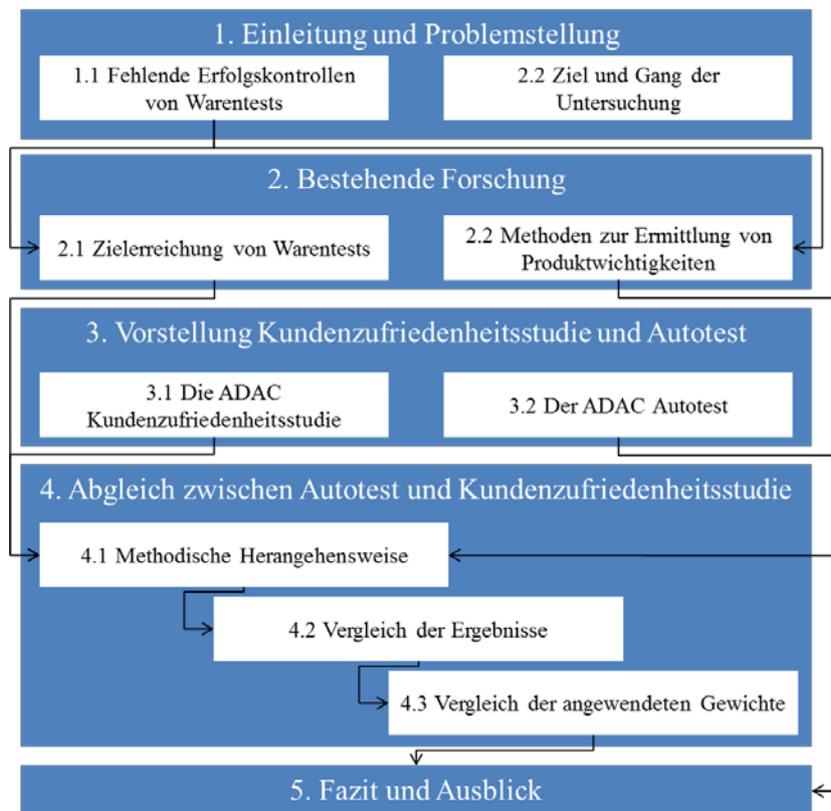


Abbildung 12: Aufbau der Untersuchung (Pfeile zeigen direkte Abhängigkeiten der Kapitel)

2 Bestehende Forschung zur Erfolgskontrolle von Warentests und Gewichtung von Produktattributen

2.1 Untersuchungen zur Zielerreichung von Warentests

Eine Möglichkeit, die Güte von Warentests zu bestimmen, ist der Vergleich mit Ergebnissen aus Kundenzufriedenheitsstudien. Dies ist möglich, da die Kundenzufriedenheit ein Konstrukt aus Erwartungen auf der einen und erlebter Leistung durch die Verbraucher auf der anderen Seite ist (Beutin, 2008: S. 813). Untersuchungen zum Zusammenhang zwischen Zufriedenheit mit Produkten und Warentestergebnissen gibt es insgesamt nur wenige. Die existierenden Studien werden nachfolgend vorgestellt und kritisch gewürdigt

Silberer et al. (1984) untersuchen den Einfluss von Warentests auf die Zufriedenheit mit dem Produkt. Dazu befragten sie 615 Verbraucher, die vor dem Kauf den vergleichenden Warentest gelesen hatten, nach ihrer Zufriedenheit und der Wiederkaufsabsicht des gewählten Produkts. Als Kontrollgruppe dienten 702 Personen, die keinen Test gelesen hatten. Zufriedenheit und Wiederkaufsabsicht unterscheiden sich dabei zwischen beiden Gruppen nicht signifikant. Der Nutzen der Testinformation wird von den Personen, die den Test vor dem Kauf gelesen hatten, jedoch als hoch eingestuft (Silberer et al. 1984: S. 121, 149–152).

Silberer et al. (1984: S. 148) beziehen sich in ihrer Untersuchung auch auf die Studien von Thorelli, Becker und Engledow (1975), Tölle, Hofacker und Kaas (1981) und Grabicke et al. (1980). Diese werden nachfolgend kurz vorgestellt.

Thorelli et al. (1975) untersuchen Unterschiede zwischen deutschen und amerikanischen Käufern. Dabei vergleichen sie auch die Zufriedenheit mit Kaufentscheidungen bei Autos. In der deutschen Studie sind die Nutzer von Warentestinformationen (70 Befragte) signifikant zufriedener als die 108 Nicht-Nutzer (S. 122-123). In der amerikanischen Studie mit 99 Lesern der Zeitschrift „*Consumer Report*“⁶ und 171 Nicht-Lesern konnte kein signifikanter Unterschied festgestellt werden (S. 86). Welche Fahrzeuge von den beiden Gruppen genau gekauft wurden, wurde jedoch nicht ausgewertet.

Tölle et al. (1981) untersuchen den Unterschied zwischen Information-Seekers (IS) und gewöhnlichen Verbrauchern. Als IS werden Personen bezeichnet, die sich vor dem Warenkauf intensiv über mögliche Produkte informieren.⁷ Befragt wurden jeweils 54 Personen. Beide Gruppen sind in ihrer sozio-ökonomischen Zusammensetzung identisch. Eines der Untersuchungsergebnisse ist, dass IS im Durchschnitt etwas weniger zufrieden mit den Produkten sind. Die Autoren führen dies jedoch auf eine höhere Anspruchshaltung zurück (S. 47-49).

Grabicke et al. (1980) untersuchen den Einfluss von Testberichten auf die Zufriedenheit beim Produktkauf. Dazu wurde 81 Probanden vor dem Kauf eines Produkts der entsprechende Testbericht der Stiftung Warentest zur Verfügung gestellt. Als Kontrollgruppe fungierten 81 Personen, die sich vorher nicht mit den Testergebnissen beschäftigt hatten (Grabicke et al., 1980: S. 246–247). Kurz nach dem Kauf war die Zufriedenheit beider Gruppen gleich hoch. Nach drei bis sechs Monaten sind die Nutzer der Testberichte jedoch zufriedener als die Nicht-Nutzer. Im Durchschnitt sind die Produkte im Test, die von den Testergebnis-Nutzern gekauft wurden, allerdings auch besser bewertet worden (Grabicke et al., 1980: S. 249–251).

Die Ergebnisse der Studien zum Einfluss von Warentests auf die Zufriedenheit sind insgesamt uneinheitlich. Weitere Studien, insbesondere auch aktuelle Studien, zum Zusammenhang zwischen Warentests und Zufriedenheit konnten im Rahmen der Recherche nicht gefunden werden. Die Untersuchungen sind in der folgenden Tabelle 4 zusammengefasst.

Autoren	Untersuchungsgegenstand	Befragte	Ergebnisse
Silberer et al. (1984)	Einfluss von Warentests auf die Produktzufriedenheit	615 Test-Nutzer vs. 702 Nicht-Nutzer	Kein signifikanter Unterschied bei Zufriedenheit und Wiederkaufsabsicht.
Thorelli et al. (1975)	Zufriedenheit deutscher und amerikanischer Käufer mit Fahrzeugen	Deutsche Studie: 70 Nutzer von Warentest Infos vs. 108 Nicht-Nutzer Amerikanische Studie: 99 Nutzer vs. 171	In der deutschen Studie sind Warentest-Nutzer zufriedener, in der amerikanischen Studie sind Nutzer nicht zufriedener.

⁶ Vergleichbar mit der Zeitschrift „*test*“ der Stiftung Warentest.

⁷ In der Untersuchung wurden Information-Seekers durch ein Abonnement der Zeitung „*test*“ der Stiftung Warentest operationalisiert.

		Nicht-Nutzer	
Tölle et al. (1981)	Unterschiede zwischen Information-Seekers (IS) und gewöhnlichen Verbrauchern	54 IS (Leser der Zeitung "test") vs. 54 Nicht-Leser. Identisch in der sozio-ökonomischen Zusammensetzung	IS sind etwas weniger Zufrieden mit den Produkten, haben nach Meinung der Autoren aber evtl. höhere Ansprüche.
Grabicke et al. (1980)	Einfluss von Warentest auf Zufriedenheit beim Produktkauf	81 Test-Nutzer vs. 81 Nicht-Nutzer	Kurz nach dem Kauf ist die Zufriedenheit gleich groß, nach drei bis sechs Monaten sind Test-Nutzer zufriedener, haben aber auch die besseren Produkte gekauft.

Tabelle 4: Bestehende Untersuchungen zwischen Warentest und Produktzufriedenheit

2.2 Untersuchungen und Methoden zur Ermittlung von Produktattributgewichtungen

Ein wesentlicher Faktor bei der Berechnung einer Gesamtnote im Warentest ist die Zuordnung der Gewichtungsfaktoren zu den einzelnen Produktattributen. Daher werden auch zu diesem Thema vorhandene Untersuchungen vorgestellt.

Silberer et al. untersuchen auf Grundlage einer Befragung, ob die Gewichtungen im Test mit den Erwartungen der Verbraucher übereinstimmen. Dazu wurden Probanden aufgefordert, die Gewichtung der einzelnen Aspekte für die Produkte „Waschmittel“ und „Farbfernseher“ festzulegen. Die Gewichtungen der Verbraucher und der Stiftung Warentest weichen dabei im Mittel voneinander ab. Die hohe Varianz der Gewichtungen der Probanden führte dazu, dass den Autoren eine Gewichts Anpassung auf Basis dieser Untersuchung nicht sinnvoll erschien, da es keine einheitliche Verbrauchererwartung gäbe. Ein weiteres Problem ist laut der Autoren, dass Verbraucher teilweise nicht in der Lage seien, Wichtigkeiten korrekt einzuschätzen, da ihnen das nötige Wissen hierzu fehle. Eine Generalisierung dieser Erkenntnisse auf andere Produktgruppen erfolgt nicht (Silberer et al., 1984: S. 206–216). Ein Problem der von Silberer et al. gewählten Erhebungsform ist, dass direkt erfragte Wichtigkeiten von Produkteigenschaften davon abhängen, inwieweit sie bei einem aktuell genutzten Produkt erfüllt werden (Swan, Combs, 1976: S. 32)⁸.

Neben diesen direkt auf Warentests bezogenen Untersuchungen zur relativen Wichtigkeit von Produkteigenschaften existiert eine Vielzahl an Methoden zur Ermittlung von Produktattributgewichtungen. Matzler und Bailom geben eine Übersicht über verschiedene Verfahren (2009: S. 288–290).

Bei der Verwendung von *Rating-Skalen* müssen Befragte die jeweilige Wichtigkeit je Attribut angeben.⁹ Ein Hauptproblem dieser Befragungsart ist, dass Befragte dazu neigen, tendenziell alle Attribute als „wichtig“ zu benennen („*Anspruchsinflation*“). Beim *Paarvergleich* werden den Befragten immer je zwei Attribute vorgelegt

⁸ So können z. B. die Schwächen des eigenen Fernsehgeräts die gefühlte Wichtigkeit einzelner Produktmerkmale beim Kauf eines neuen Geräts maßgeblich beeinflussen.

⁹ Z. B. in Form einer Zustimmungsskala von „sehr wichtig“ bis „unwichtig“

und die Befragten müssen das wichtigere Attribut wählen. Nachteil dieses Verfahrens ist, dass bei einer hohen Anzahl an Attributen die Anzahl der Paarvergleiche exponentiell steigt (Matzler, Bailom, 2009: S. 289).

Bei *Rangordnungsverfahren* müssen alle Produkteigenschaften von den Befragten nach Wichtigkeit sortiert werden. Auch diese Methode wird bei vielen Attributen schnell sehr aufwändig (Matzler, Bailom, 2009: S. 289).

Wird eine *Konstantsummenskala* verwendet, müssen die Befragten jedem Attribut Punkte zuordnen. Dabei steht jedoch nur eine konstante Punktzahl zur Verfügung, so dass sich hier relative Wichtigkeiten je befragter Person direkt aus den Antworten ergeben. All diesen Verfahren gemein ist nach Matzler und Bailom, dass die Ergebnisse schwer interpretierbar sind, da nicht bekannt ist, auf welches Qualitätsniveau sich die Wichtigkeit bezieht (siehe dazu auch Swan, Combs, 1976).¹⁰

Verfahren	Kurzbeschreibung
Rating-Skala	Direkte Abfrage der Wichtigkeit einzelner Attribute mittels Skala (z. B. "sehr wichtig" bis "unwichtig").
Paarvergleich	Befragte müssen aus jeweils zwei Attributen das für sie wichtigere auswählen. Dazu müssen alle möglichen Kombinationen gebildet und abgefragt werden.
Rangordnungsverfahren	Befragte müssen alle Attribute nach absteigender Wichtigkeit sortieren.
Konstantsummenskala	Befragte müssen auf alle Attribute eine konstante Anzahl an Punkten verteilen.
Conjoint-Analyse	Befragte müssen aus Produktvarianten ("Attributausprägungskombinationen") jeweils das bevorzugte Produkt wählen.
Regressionsanalytische Verfahren	Durch Regressionsgleichungen wird der Einfluss verschiedener Attribute auf die Gesamtzufriedenheit ermittelt. Die Regressionskoeffizienten werden als Wichtigkeiten interpretiert.

Tabelle 5: Verfahren zur Attributwichtigkeitsbestimmung

In der *Conjoint-Analyse* müssen Befragte wiederholt aus zwei (oder mehreren) Varianten eines Produktes das für sie vorteilhafteste Produkt auswählen. In jeder Wiederholung werden die Varianten in verschiedenen Attributen variiert, so dass hinterher die relative Wichtigkeit der Attribute berechnet werden kann. Die Conjoint-Analyse simuliert damit die Kauf-Entscheidungsphase und ist hierfür besonders geeignet. Sie erfordert von Konsumenten eine Abwägung zwischen unterschiedlichen Produktattributausprägungen. Auch die Conjoint-Analyse ist praktisch nur für eine sehr begrenzte Anzahl an Produkteigenschaften einsetzbar (Baier, Bruschi, 2009: S. 10). *Regressionsanalysen* erlauben es, aus Teil-Zufriedenheiten Einflussstärken auf die Gesamtzufriedenheit zu berechnen. Aus den Regressionskoeffizienten können damit Wichtigkeiten abgeleitet werden (Matzler, Bailom, 2009: S. 289–290). Alle bisher genannten Verfahren sind in Tabelle 5 zusammengefasst.

¹⁰ Bei einem Auto ist die Sicherheit ein wichtiges Kriterium, wenn sie nicht ausreichend ist, aber kein Auswahlgrund, wenn sie bei allen infrage kommenden Produkten hinreichend hoch ist (Matzler, Bailom, 2009: S. 289).

Ein Problem der bisher vorgestellten Verfahren ist die Nicht-Berücksichtigung der Tatsache, dass die relative Attributwichtigkeit häufig von der Attribut-Performance abhängt (Matzler et al., 2004: S. 272). Conjoint-Analyse und regressionsanalytische Verfahren berücksichtigen dies in der Erhebung. Durch das Ausweisen eines einzelnen Gewichtungswertes je Attribut wird diese Information in der Regel jedoch nicht ausgewertet.

Beispiele für Verfahren, welche die Attributwichtigkeit in Beziehung zur Attribut-Performance setzen, sind z. B. das von Harvey Thompson entwickelte und von Vavra vorgestellte *Importance Grid* (Vavra, 1997: S. 383–385) und die *Penalty-Reward-Contrast-Analyse* (Brandt, 1987: S. 62–64). Beide Methoden haben zum Ziel, die Attribute in Basis-, Leistungs- und Begeisterungsfaktoren einzuteilen, eine genaue Gewichtsermittlung ist nicht das primäre Ziel. Damit beziehen sich diese Methoden auf die Einteilung der Attribute nach dem *Kano-Modell* (Kano, 1984). Als Basis-Faktoren werden Attribute bezeichnet, die ein gewisses Mindestmaß erfüllen müssen, um keine Unzufriedenheit zu erzeugen. Eine höhere Attribut-Performance solcher Faktoren führt aber nicht zu höherer Zufriedenheit. Begeisterungsfaktoren werden vom Kunden nicht erwartet und führen bei Vorhandensein zu besonders hoher Zufriedenheit, ein Fehlen führt jedoch nicht zur Unzufriedenheit. Leistungsfaktoren sind eine Kombination der beiden erstgenannten Faktoren. Bei schlechter Leistung führen sie zu Unzufriedenheit, bei guter Leistung zu Zufriedenheit (Matzler, Bailom, 2009: S. 290–291).

Der Vorteil dieser Einteilung ist, dass gerade in der Produktentwicklung die verschiedenen Attribute möglichst genau den Kundenanforderungen entsprechend entwickelt werden können. Diese Einteilung hat für Warentests jedoch den Nachteil, dass die Anwendung des Verfahrens der breiten Öffentlichkeit nur schwer erklärbar ist. Da sich die aktuelle Warentest-Praxis auf einen Gewichtungswert je Attribut konzentriert, beschränkt sich die weitere Untersuchung auf Verfahren, die eindeutige Gewichte liefern.

3 Vorstellung der Kundenzufriedenheitsstudie und des Autotests

3.1 Die ADAC-Kundenzufriedenheitsstudie

Im Jahr 2012 führte der ADAC erstmalig seine Kundenzufriedenheitsstudie „*ADAC Kundenbarometer*“ durch. In der Studie, die von einem der Autoren dieser Untersuchung betreut wird, werden jährlich etwa 20.000 Personen mit jungen Fahrzeugen¹¹ nach ihrer Zufriedenheit mit ihrem Fahrzeug sowie dem Service des Hersteller befragt (ADAC, 2012c).

Die Befragung erfolgt über einen Online-Fragebogen, der ungefähr 140 Fragen umfasst und dessen Beantwortung im Durchschnitt 20 Minuten dauert. Die Grundgesamtheit der Befragung sind alle Autofahrer aus Deutschland, die über ein entsprechendes Fahrzeug und einen Internetzugang verfügen. Die Erhebung erfolgt dabei bestandsrepräsentativ: Modelle mit hohen Zulassungszahlen im genannten Erstzulassungs-

¹¹ Hierunter fallen Fahrzeuge deren Erstzulassung zwischen vier Monaten und dreieinhalb Jahren zurückliegt.

sungszeitraum sind entsprechend stärker in der Untersuchung vertreten als Nischenmodelle.

Nach der Identifikation des Fahrzeugs (siehe hierzu Schreier, Biethahn, 2012) wird zuerst die Gesamtzufriedenheit mit dem Fahrzeug abgefragt, bevor die einzelnen Teilbereiche abgefragt werden. Für jeden Teilbereich gibt es wiederum Detailfragen. Die Teilbereiche und die Detailfragen sind bewusst an den ADAC-Autotest (siehe nächstes Kapitel) angelehnt, um einen Abgleich zwischen Zufriedenheitsstudie und Autotest zu ermöglichen. Im Rahmen der Fragebogenentwicklung wurde darauf geachtet, dass Aspekte, die nur von Experten beantwortet werden können, aus dem Fragebogen gestrichen wurden, um die Erzeugung von Artefakten¹² zu vermeiden.

Im Rahmen dieser Untersuchung werden in den folgenden Kapiteln die Fragen, die zur Analyse verwendet wurden, jeweils vollständig benannt. Die Nummer vor jeder Frage gibt Aufschluss über die Position der Frage im Fragebogen an.

3.2 *Der ADAC-Autotest*

Der ADAC führte Autotests erstmals im Jahr 1926 durch (Kleinschmidt, 2010: S. 113). Seither haben sich die Entwicklung des Testvorgehens und die Gewichtungen der einzelnen Unterkriterien stark verändert. Bei der Weiterentwicklung des Autotests ähnelt das Verfahren des ADAC dem allgemeinen Vorgehen der Stiftung Warentest (vgl. Hüttenrauch, 1986: S. 18; Stiftung Warentest, 2011a: S. 55) mit dem Unterschied, dass das Expertengremium nur aus ADAC-Mitarbeitern besteht. Eine Überprüfung auf Basis empirischer Untersuchungen und Anpassung der Gewichte fand lediglich in Form einer unveröffentlichten Masterarbeit statt (Ruhdorfer, 2008).

Aktuell werden durch den ADAC jedes Jahr ca. 150 Fahrzeuge einem Autotest unterzogen und pro Fahrzeug etwa 350 Prüfpunkte ermittelt. Soweit es möglich ist, werden objektiv messbare Kriterien verwendet. Die Benotung erfolgt teilweise fahrzeugklassen-bezogen¹³ und reicht „von 0,6 (beste Note) bis 5,5 (schlechteste Note)“ (ADAC, 2013a: S. 1–2).

Die Anzahl der getesteten Fahrzeugvarianten je Modell hängt von der Marktrelevanz der Fahrzeuge in Deutschland ab. Ein Modell mit (zu erwartender) hoher Zulassungszahl wird entsprechend häufiger getestet als ein Nischenmodell mit nur wenigen verkauften Fahrzeugen pro Jahr. Weitere Kriterien für Tests sind die Neuerscheinung von Fahrzeugen (in Form neuer Modellgenerationen oder Facelifts), die Anzahl existierender Varianten, Verfügbarkeit der Fahrzeuge und die Zugriffszahlen auf die Autotests im Internet (ADAC, 2013a: S. 1). Abbildung 13 zeigt die zehn am Häufigsten getesteten Modelle im Verhältnis zur Gesamtzahl der Tests in den Jahren 2006 bis 2012. Der VW Golf wurde dabei insgesamt 28-mal getestet, während in der Zeit ins-

¹² Artefakte sind Pseudoergebnisse, die z. B. durch Überforderung der Befragten entstehen können. Wenn Probanden Fragen gestellt bekommen, deren Inhalt sie nicht verstehen oder deren Beantwortung ihr Wissen übersteigt, werden zwar Daten erfasst, diese sagen jedoch nichts über den tatsächlichen Sachverhalt aus und können zu falschen Schlüssen führen (Weinreich, von Lindern, 2008: S. 162).

¹³ Z. B. bei der Größe des Kofferraums.

gesamt 1.098 Fahrzeuge getestet wurden. Die Häufigkeit der Tests entspricht dabei im Wesentlichen den Zulassungszahlen des Zeitraums.

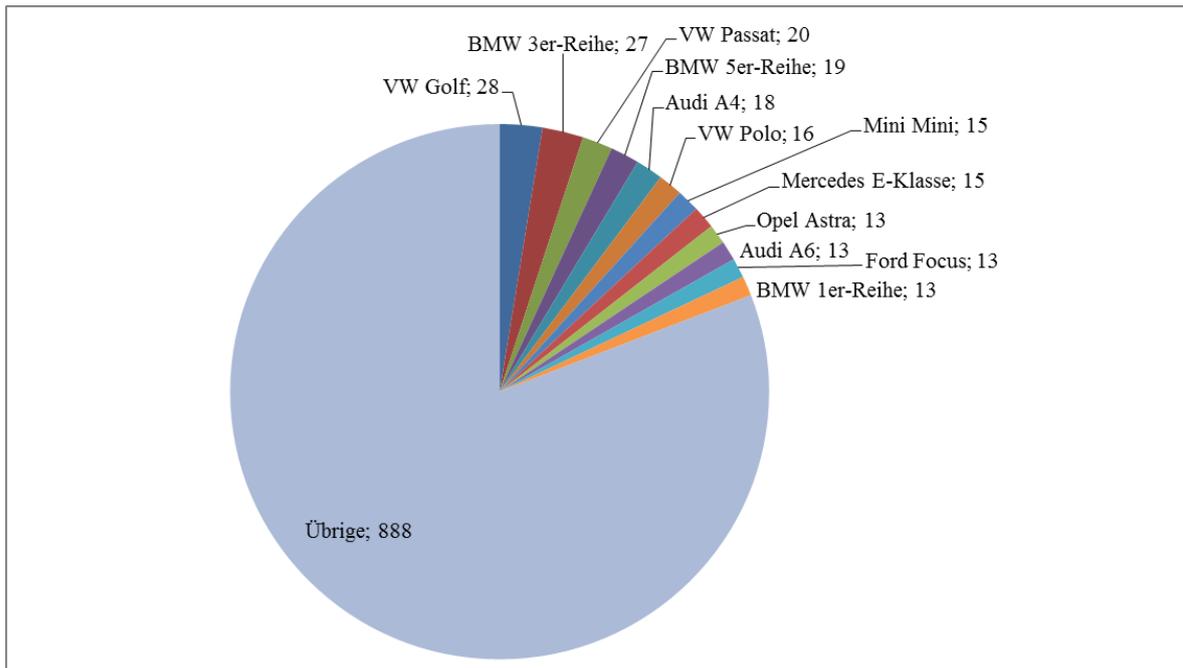


Abbildung 13: Anzahl Autotests zwischen 2006 und 2012, Top 10 Modelle und Tests insgesamt (eigene Auswertung)

Im ADAC-Autotest werden Kosten und Wirtschaftlichkeit separat im „ADAC-Urteil Wirtschaftlichkeit“ bzw. im „ADAC-Urteil Autokosten“ bewertet. Die fahrzeugbezogenen Werte werden im „ADAC-Urteil Autotest“ bzw. „ADAC-Urteil Technik / Umwelt“ zusammengefasst. Die Namen werden jeweils synonym in Veröffentlichungen des ADAC verwendet. In dieser vorliegenden Untersuchung wird mit der „Gesamtnote“ immer auf das „ADAC-Urteil Autotest“ verwiesen und mit der Note „Autokosten“ auf das entsprechende ADAC-Urteil.

Die Gesamtnote besteht aus folgenden Hauptkriterien:

- Karosserie / Kofferraum
- Innenraum
- Komfort
- Motor / Antrieb
- Fahreigenschaften
- Sicherheit
- Umwelt / Ecotest

Die beiden letztgenannten Hauptkriterien gehen mit doppelter Gewichtung in die Gesamtnote ein, die übrigen Werte jeweils mit einfacher Gewichtung. Unter diesen Hauptkriterien gibt es Unterkriterien, die immer mehrere Einzelkriterien zusammenfassen (ADAC, 2013a: S. 1–10). Über alle Ebenen erfolgt die Ermittlung der übergeordneten Note auf Basis von Gewichtungsfaktoren. Eine Ausnahme hiervon bilden die sogenannten Durchschlageffekte, die es für sicherheitsrelevante Mängel gibt. So kann

ein Fahrzeug ohne ESP¹⁴ im Unterkriterium „*Aktive Sicherheit*“ bestenfalls die Note 4,6 erreichen. Ein Modell, das im Hauptkriterium „*Sicherheit*“ die Note 4,6 oder schlechter erhält, in der Gesamtnote maximal 0,5 Punkte besser sein als im Hauptkriterium „*Sicherheit*“.

In der Bewertung der Autokosten werden Eigenschaften wie der Anschaffungspreis, die Unterhaltskosten (Versicherung, Steuern), kalkulatorische Werkstattkosten, die prognostizierte Wertstabilität sowie die Betriebskosten auf Basis des gemessenen Verbrauchs und einer angenommenen Jahresfahrleistung von 15.000 km / Jahr integriert (ADAC, 2013a: S. 11).

Jede Anpassung des Testverfahrens oder der Gewichtungen führt dazu, dass ein Vergleich von Ergebnissen aus vorherigen Tests nicht oder nur eingeschränkt möglich ist. Ein Vergleich von Produkten ist aber der Hauptzweck eines Warentests. Das ist insbesondere bei „*Continuous Tests*“, also Tests, bei denen regelmäßig neue Produkte nach einem bestehenden Verfahren getestet werden (Lösenbeck, 2003: S. 105), problematisch. Der ADAC-Autotest fällt in diese Kategorie. Hier muss ein Mittelweg zwischen der Aktualität des Testverfahrens und der Vergleichbarkeit von neuen und alten Tests gefunden werden (vgl. Lösenbeck, 2003: S. 104).

Der ADAC-Autotest wurde Anfang des Jahres 2006 weitreichend überarbeitet. Die nächste größere Überarbeitung fand im Januar 2009 statt, bevor im Rahmen der Aktualisierung des ADAC-Ecotests¹⁵ im April 2012 (ADAC, 2013b: S. 1) auch der Autotest erneut stärker überarbeitet wurde¹⁶. Zwischen diesen größeren Änderungen fanden lediglich kleinere Anpassungen statt, die jedoch nur unwesentlichen Einfluss auf die Gesamtergebnisse der getesteten Fahrzeuge hatten.

Die Anpassung des Tests im April 2012 führte zu einem Anstieg der Noten, wie aus nachfolgender Abbildung 14 ersichtlich wird. Die Ergebnisse im Jahr 2012 vor der Test-Anpassung sind dem Jahr „2012“ zugeordnet, die Werte ab April dem Jahr „2012,5“. In den Jahren von 2006 bis 2012 sind die Noten im Test im Durchschnitt etwas besser geworden. Die Anpassungen des Tests Anfang 2009 sind ebenfalls durch einen leichten Anstieg im Notenmittelwert sichtbar.

¹⁴ Elektronisches Stabilitätsprogramm - Eine Funktion, die verhindert, dass der Fahrer eines Fahrzeugs bei zu schneller Kurvenfahrt die Fahrspur verlässt, indem einzelne Reifen automatisch abgebremst werden.

¹⁵ Im ADAC-Ecotest werden Schadstoffausstoß und Kraftstoffverbrauch bewertet.

¹⁶ Neu eingeführt wurde der Aspekt „Alltagstauglichkeit“ mit Eigenschaften wie „Reichweite“ und „Tatsächliche Zuladung (in kg)“ um neue Fahrzeugkonzepte wie Elektroautos vollständig bewerten zu können. Zusätzlich wurden Zielgruppenbewertungen für bestimmte Nutzergruppen eingeführt.

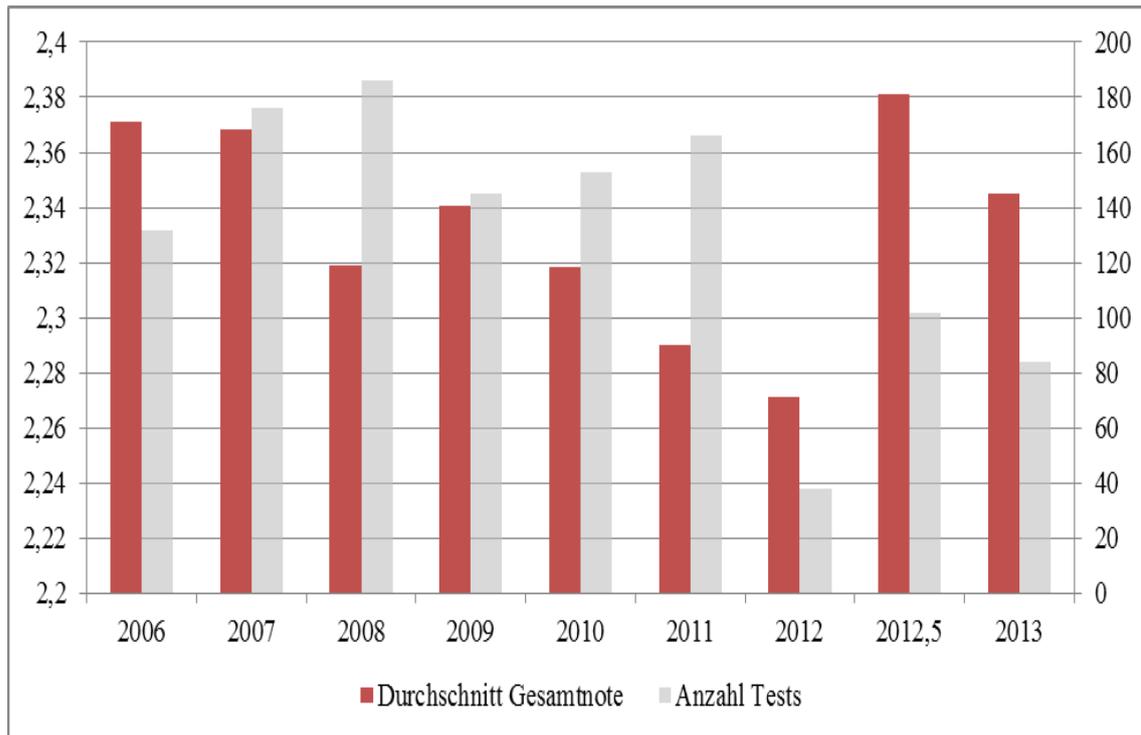


Abbildung 14: Durchschnitt Gesamtnote und Anzahl Tests nach Jahr (eigene Auswertung)

Die Neuerung in 2009 mit dem stärksten Effekt auf die Gesamtnote ist die Einführung der Durchschlageffekte, insbesondere bei fehlendem ESP oder anderen Sicherheitsmängeln. Hiervon sind wenige Fahrzeuge betroffen, die dann aber eine deutlich schlechtere Gesamtnote erhalten haben. Der Median der Gesamtnote beträgt für alle aufgelisteten Jahre 2,3.¹⁷ Der Einfluss dieser Anpassungen auf die Testergebnisse wird in den sich anschließenden Auswertungen berücksichtigt.

Die Mittelwerte der Hauptkriterien sind in Abbildung 15 dargestellt. Die Note für Wirtschaftlichkeit ist im Durchschnitt deutlich schlechter als die Noten der übrigen Hauptkriterien, welche in die Gesamtnote eingehen. Auch auf dieser Aggregationsebene sind die erhöhten Anforderungen des Ecotests deutlich sichtbar. Die Ergebnisse im Jahr 2012,5 (nach der Testanpassung) und 2013 sind deutlich schlechter als in den Jahren zuvor. Deutlich verbessert über den Zeitraum haben sich die Sicherheitsbewertung und auch die Noten für den Innenraum. Die Noten der Hauptkategorie Karosserie / Kofferraum verschlechtern sich etwas. Die übrigen Hauptkriterien bewegen sich uneinheitlich.

¹⁷ Ausnahme bildet das geteilte Jahr 2012. In den ersten drei Monaten ist der Median 2,2. Im Rest des Jahres ist er 2,4. Für das Gesamtjahr ist der Median auch 2,3.

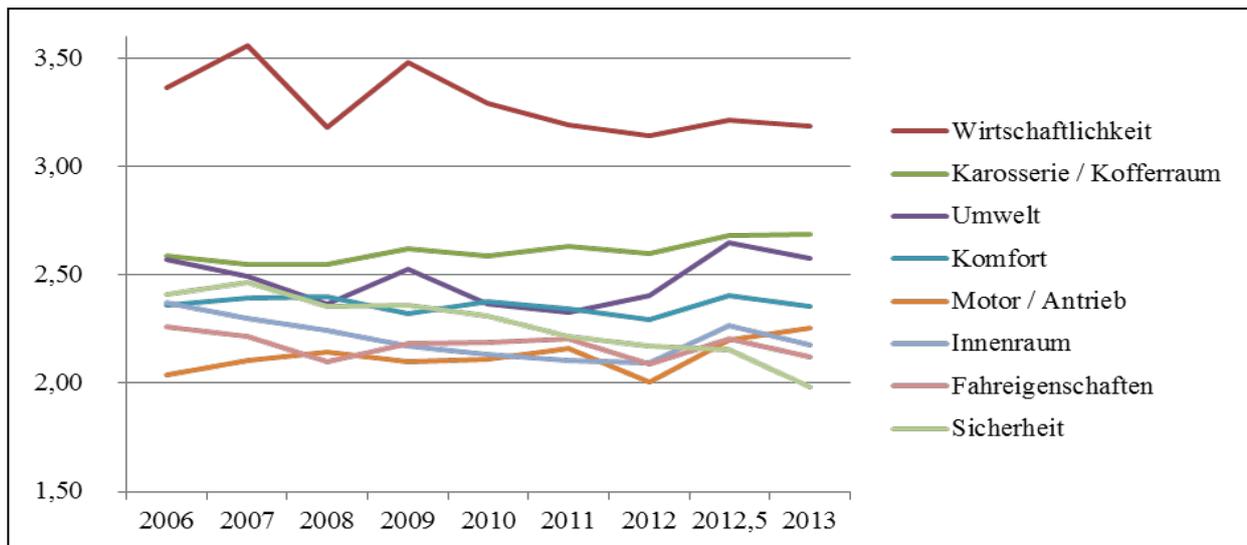


Abbildung 15: Mittelwerte der Hauptkriterien (inkl. Wirtschaftlichkeit) pro Jahr (eigene Auswertung)

4 Abgleich zwischen Autotest und Kundenzufriedenheitsstudie

4.1 Methodische Herangehensweise

Ziel von Warentests ist es, Verbrauchern bei der Auswahl für das für sie beste Produkt behilflich zu sein (vgl. Stiftung Warentest, 2011b § 2). Dazu werden die wesentlichen Kriterien festgelegt und in Gesamt- und Teilnoten bewertet (Hüttenrauch, 1986: S. 21). Für diese beiden Ebenen¹⁸ wird nun eine Methode zum Abgleich zwischen den Ergebnissen des Autotests und der Kundenzufriedenheitsstudie hergeleitet.

Das Ziel einer Gesamtnote in Tests ist die Verdichtung aller Informationen auf einen Wert. Dies soll die Informationsgewinnung für den Verbraucher vereinfachen und für eine hohe Testverbreitung durch andere Medien sorgen. Erstens greifen nicht alle Personen auf die Testinformationen zurück, zweitens entscheiden sich nicht alle Personen, die auf die Testinformation zurückgegriffen haben, für das Produkt, welches im Test am besten abgeschnitten hat. Dennoch sollte das am besten bewertete Produkt im Durchschnitt die höchste Gesamtzufriedenheit hervorrufen. Analog sollte das zweitbeste Produkt den zweitbesten Zufriedenheitswert erreichen. Dies geht so weiter, bis schließlich das am schlechtesten bewertete Produkt auch den niedrigsten Zufriedenheitswert aufweist.

Im ersten Schritt soll deshalb überprüft werden, ob die Test-Gesamtnote mit der Gesamtzufriedenheit korreliert. Dazu wird folgende Alternativhypothese formuliert:

H₁: Die Gesamtnote im Autotest weist einen signifikant messbaren Zusammenhang mit der Gesamtzufriedenheit auf.

Da die Noten der Hauptkriterien des Autotests ebenfalls veröffentlicht werden und deren Zustandekommen ebenfalls Gewichtungen unterliegt, soll auch für diese Noten

¹⁸ Der ADAC-Autotest verwendet noch zwei detaillierte Ebenen. Diese sind jedoch aus Platzgründen nicht Teil dieser Untersuchung.

die Korrelation mit dem jeweiligen Zufriedenheitswert untersucht werden. Die Alternativhypothesen werden in folgender Form gebildet:

Die Note für das Hauptkriterium ... weist einen signifikant messbaren Zusammenhang mit der Zufriedenheit für diesen Teilbereich auf.

Die Bezeichnungen der Alternativhypothesen und die entsprechenden Kriterien sind:

H_{II} : Karosserie / Kofferraum

H_{III} : Innenraum

H_{IV} : Komfort

H_V : Motor / Antrieb

H_{VI} : Fahreigenschaften

H_{VII} : Sicherheit

H_{VIII} : Wirtschaftlichkeit

Die Überprüfung des Hauptkriteriums „Umwelt / Ecotest“ ist nicht möglich, da hierzu keine Frage im Fragebogen enthalten war.

Abschließend soll überprüft werden, inwieweit die Gewichte der Gesamtnote des Autotests mit den aus der Kundenzufriedenheitsstudie berechneten Gewichten übereinstimmen. Im Autotest werden Gewichte zurzeit einheitlich für alle Performance-Levels verwendet. Der in Kapitel 0 aufgezeigte Zusammenhang zwischen Gewichtung und Attribut-Performance wird für diese Überprüfung deshalb nicht betrachtet.

4.2 Vergleich der Ergebnisse

Die Daten aus der Zufriedenheitsstudie beziehen sich auf Fahrzeuge, deren Erstzulassung zum Zeitpunkt der Befragung zwischen vier Monaten und dreieinhalb Jahren zurück lag (ADAC, 2012c). Die Erstzulassung wird als Ersatz für den von Laien nur schwer ermittelbaren Produktionszeitpunkt herangezogen. Die Befragung fand im April / Mai 2012 statt, entsprechend wurden Fahrzeuge untersucht, deren Erstzulassung zwischen Oktober 2008 und Dezember 2011 lag. Für die Auswahl der Autotests bedeutet das, dass nur Fahrzeuge verglichen werden können, die während eines Großteils dieses Zeitraums produziert und verkauft worden sind. Fahrzeuge, deren Produktionsstart nach Juni 2011 bzw. deren Produktionsende vor Juli 2008 lag, werden deshalb nicht berücksichtigt. Die Beschränkung des Testzeitraums wurde so gewählt, dass alle Fahrzeuge, die zwischen 2006 und dem 3. Mai 2013 getestet wurden, berücksichtigt werden. Der Zusammenhang der Zeiträume ist in Abbildung 16 dargestellt.

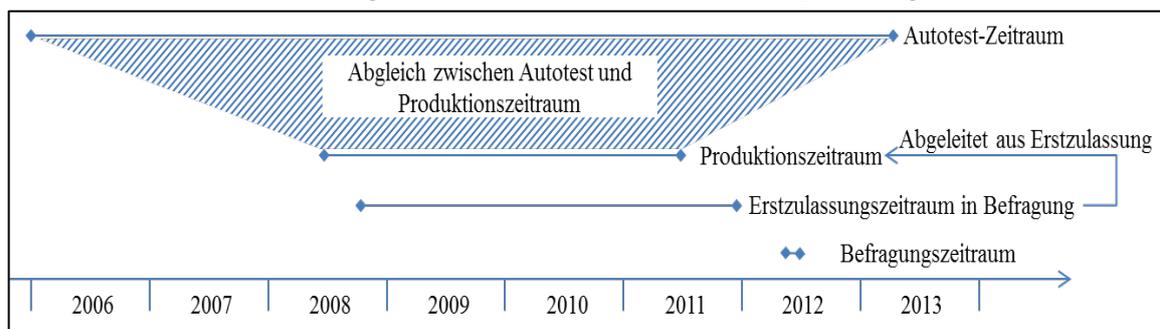


Abbildung 16: Zusammenhang der Zeiträume von Autotest und Zufriedenheitsbefragung

Um eine zufällige Schwankung zu minimieren, werden nur diejenigen Modelle verglichen, bei denen mindestens 50 Interview-Antworten vorliegen. Wendet man diese Kriterien an, bleiben insgesamt 550 Autotests mit 88 Modellen und 16.569 Interviews der Zufriedenheitsstudie für den Vergleich übrig.

Damit der Vergleich von Autotest-Noten und Zufriedenheitsstudie durchgeführt werden kann, muss sichergestellt sein, dass die Ergebnisse auf der richtigen Aggregationsebene verglichen werden. Das ist dann der Fall, wenn die Standardabweichung der Autotest-Noten des gleichen Modells in unterschiedlichen Varianten gering ist. Dazu wurden die Gesamtnote und die Noten der Hauptkriterien zuerst auf Modellebene verglichen.

Der Mittelwert der Standardabweichung der 88 Modelle beträgt 0,14 mit einem Maximum von 0,78. Nur acht Modelle haben eine Standardabweichung von mehr als 0,20. Davon ist bei sieben Modellen das fehlende bzw. nur gegen Aufpreis erhältliche ESP die wesentliche Ursache. Die Spannweite zwischen der besten und schlechtesten Gesamtnote liegt bei 50 % der 88 Modelle bei einem Wert bis 0,2. 87,5 % der Modelle haben eine Spannweite bis 0,4 (siehe Abbildung 17). Insgesamt unterscheiden sich die Noten auf Modellebene so wenig, dass der Abgleich zur Gesamtnote auf dieser Ebene vorgenommen wird.

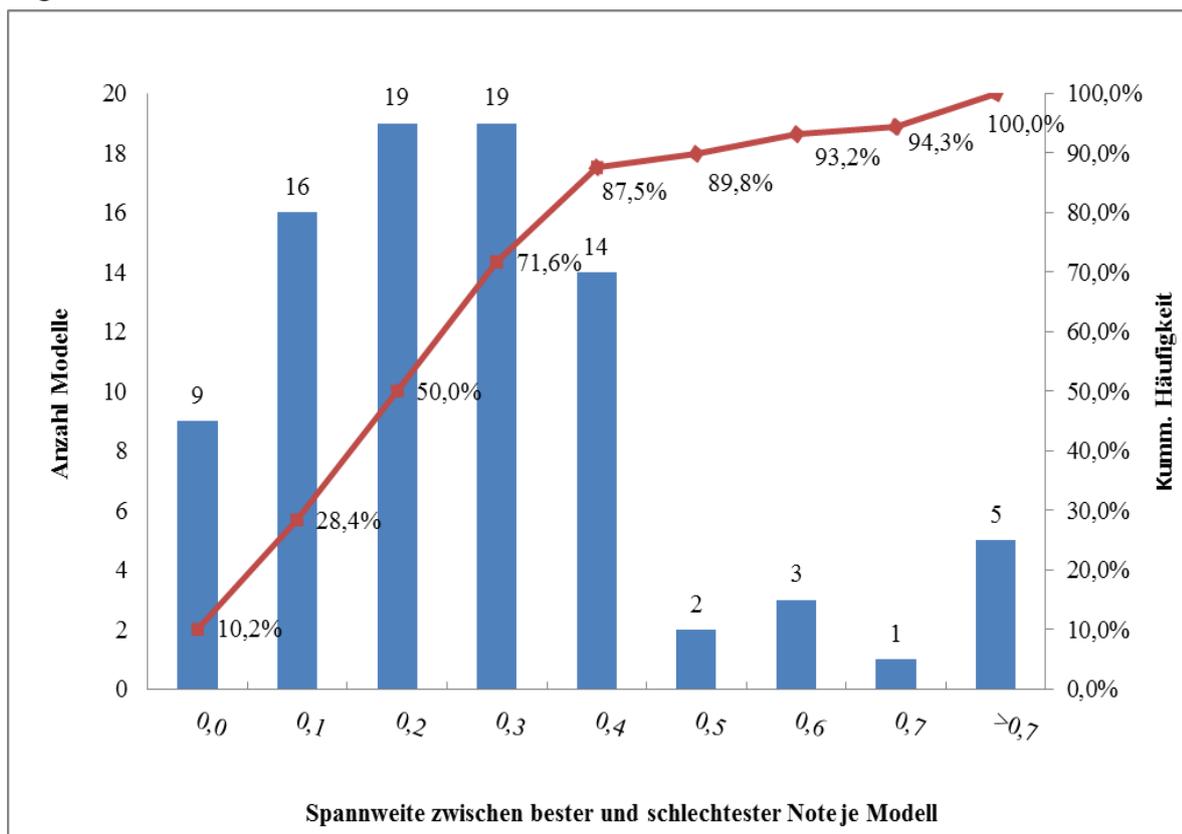


Abbildung 17: Häufigkeitsverteilung der Spannweiten zwischen bestem und schlechtestem Testergebnis je Modell (Lesebeispiel: Der Abstand 0,1 kommt 16-mal vor)

Aus der Zufriedenheitsstudie wird die zwölfte Frage „Wie beurteilen Sie Ihr Fahrzeug insgesamt“ verwendet, die über eine linksschiefe, fünfstufige Antwortskala¹⁹ verfügt. Die Antworten der Skala werden mit Werten von 1 = Hervorragend bis 5 = Schlecht bewertet. Die Ergebnisse des Abgleichs sind in Abbildung 18 dargestellt.

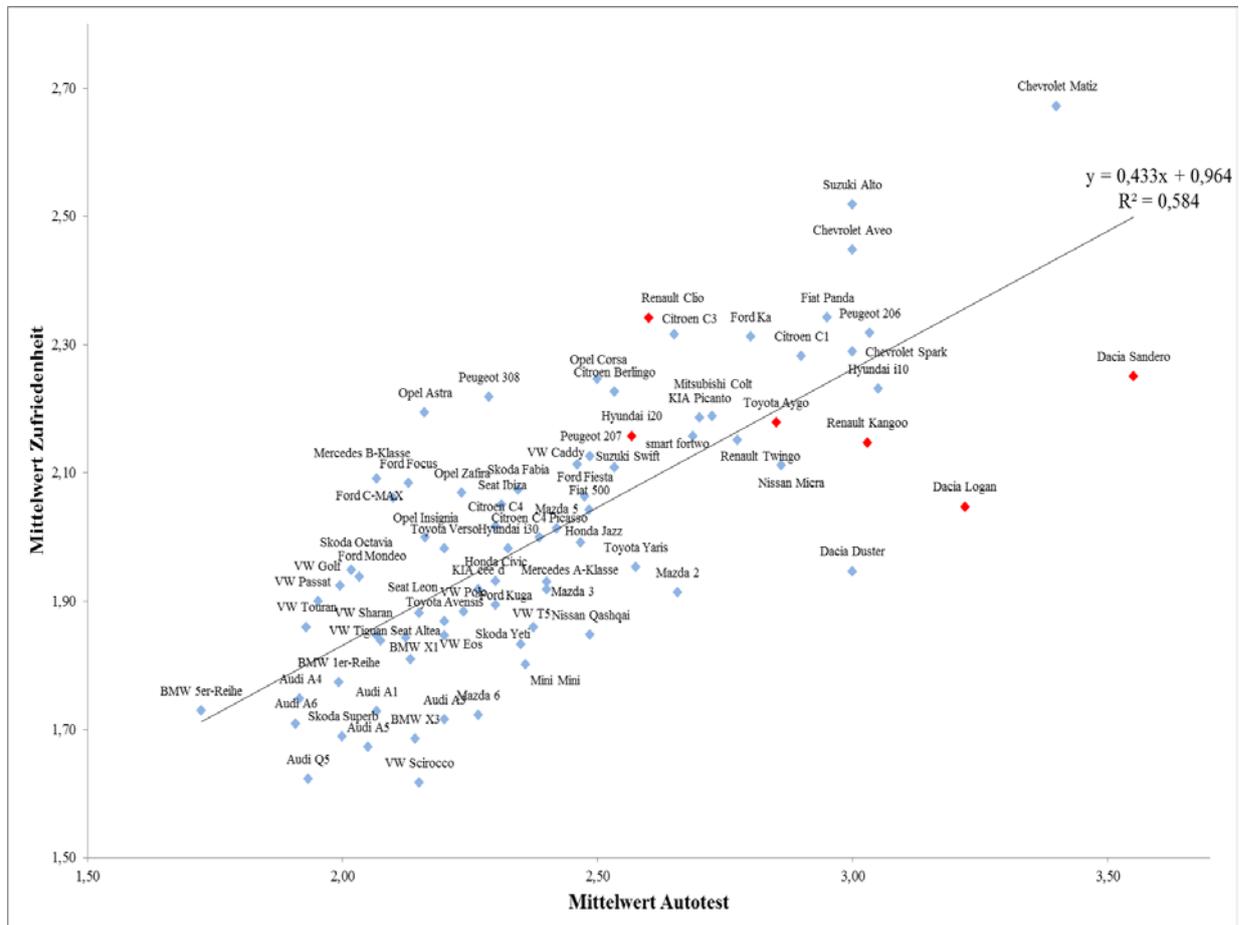


Abbildung 18: Mittelwerte von Zufriedenheit und ADAC-Autotest (Ausschnitt; rot markierte Modelle verfügen über eine hohe Standardabweichung der Autotest Gesamtnote)

Die vorher identifizierten Modelle mit hoher Standardabweichung zwischen den einzelnen Autotests sind in der obigen Abbildung rot markiert. Eine eindeutige Abweichung in eine bestimmte Richtung ist für diese Modelle nicht festzustellen. Auffällig ist das starke Abweichen der Fahrzeuge der Marke Dacia. Diese weisen höhere Zufriedenheitswerte auf, als es die eingezeichnete Regressionsgleichung erwarten ließe. Dies lässt sich durch die Nicht-Berücksichtigung der Preis-Komponente in der Autotest-Gesamtnote erklären, da bei Dacia-Fahrern der Kaufpreis einer der Haupttreiber der Zufriedenheit ist. Die ermittelte Regressionsgleichung für die vorliegenden Daten lautet:

$$\text{Gesamtzufriedenheit} = 0,964 + 0,433 * \text{Gesamtnote}_{\text{Autotest}}$$

Formel 1: Regressionsgleichung Gesamtzufriedenheit mit Mittelwert Autotest-Gesamtnote

¹⁹ Die Skala verfügt über die Antworten: Hervorragend, Sehr gut, Gut, Annehmbar und Schlecht. Zusätzlich gibt es noch die Option „Weiß nicht“.

Der F-Test für dieses Regressionsmodell hat einen Wert von 120,603 (siehe Tabelle 6). Bei 1 und 86 Freiheitsgraden bedeutet dies einen hoch signifikanten Zusammenhang ($p < 0,001$). Die Alternativhypothese H_I kann angenommen werden, die Gesamtnote im Autotest weist einen hoch signifikanten Zusammenhang zur Zufriedenheit mit dem Fahrzeug auf. Die Erklärung der Streuung für dieses Modell ist mit 58,4 % der Gesamtstreuung bereits relativ hoch. Bezogen auf den Mittelwert der Gesamtzufriedenheit der Modelle von 2,00 beträgt der Standardfehler des Schätzers 6,9 % ($= 0,138/2,00$), was als gut bewertet werden kann (Backhaus et al., 2006: S. 73).

Modellzusammenfassung									
Modell	R	R ²	Korrigiertes R-Quadrat	Standardfehler des Schätzers	Änderungsstatistiken				
					Änderung in R ²	Änderung in F	df 1	df 2	Sig. Änderung in F
1	0,764 ^a	0,584	0,579	0,13802	0,584	120,603	1	86	,000

a. Einflussvariablen : (Konstante), Mittelwert Autotest

Tabelle 6: Modellzusammenfassung zur Regressionsgleichung Gesamtzufriedenheit mit dem Mittelwert der Autotest-Gesamtnote

Zur Überprüfung der Hypothesen H_{II} bis H_{VIII} werden zuerst die Standardabweichungen der Testergebnisse je Modell untersucht (siehe Tabelle 7). Die höchste Abweichung weist die Wirtschaftlichkeit mit 0,67 auf. Die Hauptursache hierfür sind stark motorisierte Modelle, die in der Wirtschaftlichkeitsbetrachtung schlechte Noten bekommen. In der Kategorie „Motor / Antrieb“ sind die Schwankungen nicht durch eine bestimmte Art von Fahrzeugen bzw. Motoren begründet. Im Bereich „Umwelt“ ist die hohe Abweichung im Wesentlichen durch die Anpassung der Bewertungslogik in 2012 zu erklären.

	Noten Mittelwert über Modelle	Mittelwert der Standardabweichungen über Modelle
Karosserie / Kofferraum	2,56	0,15
Innenraum	2,25	0,15
Komfort	2,48	0,17
Motor / Antrieb	2,29	0,36
Fahreigenschaften	2,40	0,24
Umwelt	2,40	0,32
Sicherheit	2,43	0,21
Wirtschaftlichkeit	3,10	0,67

Tabelle 7: Mittelwerte und Standardabweichung der Hauptkriterien über die untersuchten Modelle

Analog zur Überprüfung von H_I werden auch für H_{II} bis H_{VIII} Regressionsgleichungen aufgestellt. Ausgewählte Ergebnisse finden sich in Tabelle 8. Dabei können

grundsätzlich alle Nullhypothesen verworfen und die Alternativhypothesen angenommen werden. In allen Fällen existiert ein hoch signifikanter Zusammenhang zwischen der Note des Hauptkriteriums im Autotest und der Zufriedenheitsbewertung.

Hypothese	Hauptkriterium	R ²	Konstante der Regressionsgleichung	Koeffizient des Hauptkriteriums der Regressionsgleichung	F-Wert	Signifikanz F-Wert	Prozentuale Abweichung vom Mittelwert für den Standardfehler des Schätzers
II	Karosserie / Kofferraum	0,591	0,78	0,56	124,2	0,0000	7,7%
III	Innenraum	0,428	1,44	0,36	64,4	0,0000	8,2%
IV	Komfort	0,661	0,59	0,85	168,1	0,0000	7,9%
V	Motor / Antrieb	0,684	1,23	0,53	186,3	0,0000	8,0%
VI	Fahreigenschaften	0,635	1,34	0,37	149,6	0,0000	7,3%
VII	Sicherheit	0,573	1,06	0,51	115,5	0,0000	10,5%
VIII	Wirtschaftlichkeit	0,151	2,35	0,09	15,3	0,0002	6,1%

Tabelle 8: Ergebnisse der Regressionsanalysen für die Hypothesen II bis VIII

Die erklärten Varianzen (R²-Werte) schwanken jedoch stark. Den höchsten Wert liefert die Regression für Motor / Antrieb mit 68,4 % erklärter Varianz. Den niedrigsten Wert liefert die Regressionsanalyse der Wirtschaftlichkeit mit 15,1 %. Für Wirtschaftlichkeit und Motor / Antrieb ist die Standardabweichung der Modelle im Autotest am höchsten (vgl. Tabelle 7). Ein Ausschluss der Modelle mit hoher Standardabweichung²⁰ führte zu keiner Verbesserung der erklärten Varianz. Der Mittelwert ist für Modelle mit stark schwankenden Noten genauso gut geeignet wie für Modelle mit niedriger Standardabweichung.

4.3 Vergleich der angewendeten Gewichte

Der Zusammenhang zwischen der Autotest-Gesamtnote und der Gesamtzufriedenheit kann bereits jetzt als hoch bezeichnet werden. Demgegenüber stehen teilweise niedrige Zusammenhänge zwischen den Noten der Hauptkriterien und den jeweiligen Teilzufriedenheitswerten. Im Folgenden wird durch Anwendung von Optimierungsverfahren auf die Gewichtungen des Autotests der Zusammenhang zwischen der Gesamtnote und der Gesamtzufriedenheit erhöht.

Einige Noten der Hauptkriterien korrelieren stark.²¹ Das bedeutet, dass eine eindeutige Bestimmung der optimalen Gewichte nicht möglich ist bzw. Gewichte in der Regressionsgleichung negativ werden können, was inhaltlich jedoch schwer begründbar ist.²² Die Variable Wirtschaftlichkeit kompensiert irrelevante Varianzanteile anderer Variablen und erhöht dadurch die Gesamtaufklärung der Regressionsgleichung. Der auf diese Weise maximal erreichte R²-Wert einer linearen Regression liegt bei

²⁰ Als „hoch“ wurde größer 1 bei der Wirtschaftlichkeit bzw. größer 0,5 bei Motor / Antrieb definiert.

²¹ Die Werte schwanken zwischen 0,117 und 0,819.

²² Dabei handelt es sich um einen sogenannten „Suppressionseffekt“: Ein oder mehrere Koeffizienten erhalten ein negatives Vorzeichen und unterdrücken dadurch irrelevante Varianzanteile anderer Variablen. Dadurch steigt die Varianzaufklärung des Modells insgesamt, eine Interpretation der Koeffizienten ist jedoch nicht mehr möglich (Rudolf, Müller, 2012: S. 399).

0,652. Aufgrund der Korrelationen ergeben sich allerdings viele sehr unterschiedliche Kombinationen der Gewichtungen, die zu diesem R^2 -Wert kommen.

Als Variante der Regressionsanalyse begrenzt die *Shapley Value Regression* dieses Problem, indem für jeden Regressionskoeffizient der marginale Einfluss ermittelt wird. Zusätzlich wird die Nebenbedingung erfüllt, dass jeder Koeffizient größer Null ist. Dazu werden für jeden Koeffizienten alle möglichen Regressionsgleichungskombinationen unter Berücksichtigung bzw. Ausschluss der anderen Koeffizienten berechnet. Der finale Wert je Koeffizient ergibt sich als gewichteter Mittelwert der verschiedenen Szenarien.²³

Durch Anwendung der Shapely Value Regression kann ein R^2 -Wert von 0,639 erreicht werden. Die Koeffizienten der Gleichung bzw. die Gewichte für den Autotest sind in nachfolgender Tabelle abgebildet.

	ADAC-Autotest	Shapely Value Regression	Delta in Prozentpunkten
Karosserie / Kofferraum	11 %	4 %	7
Innenraum	11 %	11 %	0
Komfort	11 %	20 %	-9
Motor / Antrieb	11 %	29 %	-18
Fahreigenschaften	11 %	20 %	-9
Umwelt	22 %	3 %	19
Sicherheit	22 %	13 %	9

Tabelle 9: Gewichte im ADAC-Autotest und mittels Shapley Value Regression ermittelte Gewichte

Die so ermittelten Gewichte unterscheiden sich teilweise deutlich von den aktuell im Autotest verwendeten Gewichten. Bei dem Versuch den Autotest auf seine Vorhersagekraft zur Kundenzufriedenheit zu optimieren, spielt der Umweltaspekt bei der Fahrzeugauswahl in der Mehrheit der Fälle nur eine untergeordnete Rolle. Die Leistung des Motors (im weiteren Sinne) hat dagegen einen deutlich höheren Einfluss auf die Zufriedenheit. Die Hauptkriterien Karosserie / Kofferraum und Sicherheit würden um 7 % bzw. 9 % geringer gewichtet werden, während Komfort und Fahreigenschaften ein um je 9 % höheres Gewicht erhalten müssten, um den R^2 -Wert zu maximieren. Durchschlageffekte werden bei den neuen Gewichten keine verwendet.

²³ Für eine genaue Beschreibung der Shapley Value Regression siehe (GfK, 2008).

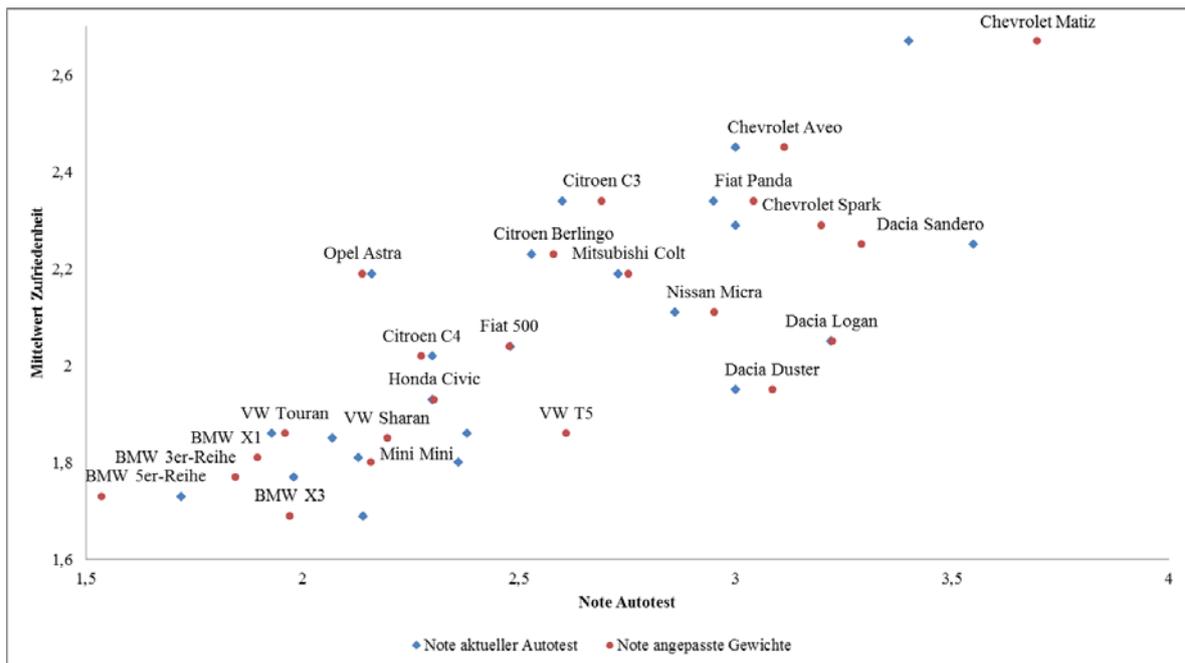


Abbildung 19: Noten Autotest mit aktueller und neuer Gewichtung (Ausschnitt)

Die unterschiedlichen Noten für einen Teil der untersuchten Modelle sind in Abbildung 18 zu sehen. Neben der Anpassung der Gewichtung wurde der Mittelwert der neuen Noten um 0,96 angehoben, um Noten auf gleichem Niveau zu erhalten. Da die Zufriedenheitswerte gleich bleiben, handelt es sich jeweils nur um Verschiebungen der Punkte auf der horizontalen Achse. Die deutlichsten Verbesserungen würden dabei die Modelle Dacia Sandero (0,26 Noten besser), BMW X1 (0,23) und Mini (0,20) erfahren. Die drei Modelle mit der stärksten Verschlechterung wären Chevrolet Matiz (-0,30), Volkswagen T-5 (-0,23) und Chevrolet Spark (-0,20). Weiterhin das Modell mit der besten Note ist die BMW 5-er Reihe. Die schlechteste Note hätte nach der neuen Berechnung nicht mehr der Dacia Sandero, sondern der Chevrolet Matiz. Die aktuellen Noten verfügen über eine Spannweite von [1,72; 3,55]. Die Noten-Spannweite der angepassten Gewichte ist [1,54; 3,69] und nutzt damit das Notenspektrum etwas besser aus.

5 Fazit und Ausblick

Die Gesamtnote des ADAC-Autotests weist bereits jetzt auf der Modellebene einen hohen Zusammenhang mit der Gesamtzufriedenheit von Autofahrern der jeweiligen Modelle auf. Durch Anpassung der Gewichte kann dieser Zusammenhang weiter erhöht werden.

Da die Ergebnisse des Autotests europaweit genutzt werden, sollten auch länderübergreifende Studien durchgeführt werden, um kulturelle Unterschiede in den Anforderungen aufzudecken und entsprechend zu berücksichtigen. Seit dem Jahr 2012 werden im Autotest auch zielgruppenspezifische Gesamtnoten ausgewiesen. Eine Überprüfung bzw. Optimierung der hierfür verwendeten Noten ist ebenfalls von Interesse.

Auf Ebene der Hauptkriterien sind die Zusammenhänge mit der Zufriedenheit existent, teilweise aber deutlich niedriger. Insbesondere in der Beurteilung der Wirt-

schaftlichkeit weichen Endkunden und Autotest stark voneinander ab. Hier existiert weiterer Analysebedarf.

Wie in Kapitel 0 dargestellt, ist die Wichtigkeit von Attributen abhängig von der Attribut-Performance. Im Falle der getesteten Autos scheint diese Eigenschaft insgesamt nur geringen Einfluss zu haben, da das Produktniveau der untersuchten Fahrzeuge hoch ist. Für die Generalisierung auf andere Produktgruppen scheint hier zusätzlicher Forschungsbedarf angebracht. Eine mögliche Lösung unter Beibehaltung der einfachen Gewichtungssparameter scheint die Einführung weiterer Durchschlageffekte. Diese bilden letztendlich nichts anderes ab als unterschiedliche Wichtigkeiten bei Nichterfüllung bestimmter Attribut-Performances.

6 Literaturverzeichnis

- ADAC (2012a): „18 Millionen ADAC Mitglieder“. Abgerufen am 17.06.2013 von http://www.adac.de/infotestrat/adac-im-einsatz/motorwelt/18_Millionen.aspx.
- ADAC (2012b): „ADAC Kundenbarometer“. Abgerufen am 06.07.2013 von <http://www.adac.de/infotestrat/autodatenbank/kundenbarometer/default.aspx?ComponentId=148354&SourcePageId=8749&quer=kundenbarometer>.
- ADAC (2012c): „ADAC Kundenbarometer: Methodik“. Abgerufen am 05.07.2013 von <http://www.adac.de/infotestrat/autodatenbank/kundenbarometer/methodik.aspx?ComponentId=149706&SourcePageId=148368>.
- ADAC (2013a): „Der ADAC-Autotest“. Abgerufen am 02.08.2013 von http://www.adac.de/_mmm/pdf/Der%20ADAC%20Autotest%20360%20KB_30012.pdf.
- ADAC (2013b): „Ecotest“. Abgerufen am 08.06.2013 von http://www.adac.de/_mmm/pdf/27473_118924.pdf.
- Backhaus, Klaus; Erichson, Bernd; Plinke, Wulff; et al. (2006): *Multivariate Analysemethoden: Eine anwendungsorientierte Einführung*. 11. Aufl. Berlin [u.a.].
- Baier, Daniel; Bruschi, Michael (2009): „Erfassung von Kundenpräferenzen für Produkte und Dienstleistungen“. In: Baier, Daniel; Bruschi, Michael (Hrsg.) *Conjointanalyse*. Berlin, S. 3–17.
- Beutin, Nikolas (2008): „Messung von Kundenzufriedenheit und Kundenbindung“. In: Herrmann, Andreas; Homburg, Christian; Klarmann, Martin (Hrsg.) *Handbuch Marktforschung: Methoden - Anwendungen - Praxisbeispiele*. 3. Auflage. Wiesbaden, S. 813–846.
- Brandt, D. Randall (1987): „A Procedure for Identifying Value-Enhancing Service Components Using Customer Satisfaction Survey Data“. In: Surprenant, C. (Hrsg.) *Add Value to Your Service*. Chicago, S. 61–65.
- Brandt, D. Randall (1988): „How service marketers can identify value-enhancing service elements“. In: *Journal of Services Marketing*. 2 (3), S. 35–41, DOI: 10.1108/eb024732.
- GfK (2008): „Getting Better Regression Results with Shapley Value Regression“. Abgerufen am 31.08.2013 von <http://marketing.gfkamerica.com/website/articles/ShapelyValueRegression.pdf>.
- Grabicke, K.; Raffée, Hans; Schätzle, T.; et al. (1980): „Die Veränderung des Informationsstandes und ihre Auswirkungen auf den Kaufentscheidungsprozess privater Haushalte“. In: Hartmann, K. D.; Köppler, K. F. (Hrsg.) *Fortschritte der Marktpsychologie*. Frankfurt/M., S. 237–257.
- Hüttenrauch, Roland (1986): „Zur Methodik des vergleichenden Warentests“. In: *Vergleichender Warentest. Testpraxis, Testwerbung, Rechtsprechung*. Landsberg am Lech, S. 13–36.

- Kano, N. (1984): „Attractive Quality and Must-be Quality“. In: *Journal of the Japanese Society for Quality Control*. Vol. 14 (No. 2), S. 39–48.
- Kleinschmidt, Christian (2010): „Comparative Consumer Product Testing in Germany“. *Business History Review*. Spring 2010 (84), S. 105–124.
- Laak, Heike van (2004): 40 Jahre Stiftung Warentest. Berlin.
- Lösenbeck, Hans-Dieter (2003): Stiftung Warentest. Ein Rückblick. 1964 - 2002. Berlin.
- Matzler, Kurt; Bailom, Franz (2009): „Messung von Kundenzufriedenheit“. In: Hinterhuber, Hans H.; Matzler, Kurt (Hrsg.) *Kundenorientierte Unternehmensführung*. 6. Aufl. Wiesbaden, S. 268–297.
- Matzler, Kurt; Bailom, Franz; Hinterhuber, Hans H.; et al. (2004): „The asymmetric relationship between attribute-level performance and overall customer satisfaction: a reconsideration of the importance–performance analysis“. In: *Industrial Marketing Management*. 33 (4), S. 271–277, DOI: 10.1016/S0019-8501(03)00055-5.
- Rudolf, Matthias; Müller, Johannes (2012): *Multivariate Verfahren: eine praxisorientierte Einführung mit Anwendungsbeispielen in SPSS*. 2. Aufl. Göttingen [u.a.]. — ISBN: 9783801724030
- Ruhdorfer, Martin (2008): Weiterentwicklung des ADAC Autotest. Universität Augsburg.
- Schreier, Jan Hendrik; Biethahn, Niels (2012): „Validität und Identifikation von Fahrzeugen und Fahrzeugeigenschaften in Kundenbefragungen“. In: *Mobility in a Globalised World*. Bamberg.
- Silberer, Günter (1985): „The impact of comparative product testing upon consumers. Selected findings of a research project“. In: *Journal of Consumer Policy*. 8 (1), S. 1–27, DOI: 10.1007/BF00380280.
- Silberer, Günter; Raffée, Hans; Fritz, Wolfgang; et al. (1984): „Die Beurteilung von Testergebnissen und Testkriterien der Stiftung Warentest durch die Verbraucher“. In: Raffée, Hans; Silberer, Günter (Hrsg.) *Warentest und Konsument: Nutzung, Wirkungen und Beurteilung des vergleichenden Warentests im Konsumentenbereich*. Frankfurt u.a., S. 187–216.
- Stiftung Warentest (2011a): „Jahresbericht 2010“. Abgerufen am 21.04.2013 von <http://www.test.de/filestore/Jahresbericht2010.pdf?path=/91/22/b4177086-a386-460c-9e8e-ebd6bba88031-file.pdf&key=2345E932F50D0B6932FDF766A8917BEC8D8B70D10>.
- Stiftung Warentest (2011b): „Satzung Stiftung Warentest“. Abgerufen am 31.05.2013 von <http://www.test.de/unternehmen/stiftungsgremien/satzung/>.
- Stiftung Warentest, Forsa (2007): „Bekanntheit der Stiftung Warentest“. Abgerufen am 22.05.2013 von <http://www.test.de/presse/bilder/symbolbilder/?ft=bild&fd=infografik-bekanntheit>.
- Swan, John E.; Combs, Linda Jones (1976): „Product Performance and Consumer Satisfaction: A New Concept“. In: *Journal of Marketing*. 40 (2), S. 25–33.
- Thorelli, Hans Birger; Becker, Helmut; Engledow, Jack (1975): *The information seekers: An international study of consumer information and advertising image*. Cambridge, Mass.
- Tölle, H.; Hofacker, T.; Kaas, K. P. (1981): „Der „Information seeker“: Konsumbegeistert oder konsumkritisch?“. In: *Marketing. Zeitschrift für Forschung und Praxis*. 3. Jg. (Heft 1), S. 47–50.
- Vavra, Terry G. (1997): *Improving your measurement of customer satisfaction*. Milwaukee, Wis. — ISBN: 0873894057
- Weinreich, Uwe; von Lindern, Eike (2008): *Praxisbuch Kundenbefragungen; Repräsentative Stichproben auswählen - Ergebnisse richtig interpretieren*. 1. Auflage. München.