Sind Rankings inhärent willkürlich?

Dominik Rohn und Karsten Weihe[1]

30. Juli 2013

Inhaltsverzeichnis

[1]*Korrespondenzadresse*: Prof. Dr. Karsten Weihe, Technische Universität Darmstadt, Fachbereich Informatik, Hochschulstraße 10, 64289 Darmstadt, weihe@cs.tu-darmstadt.de

1 Sie müssen uns kein Wort glauben

Sie können alle unsere Aussagen überprüfen. Dafür reichen die Daten, die wir in diesem kurzen Aufsatz zusammengestellt haben, völlig aus. Dass die von uns zugrunde gelegten Ausgangsdaten richtig sind, können Sie zumindest bei den von uns untersuchten Printmedien durch leicht erreichbare, öffentlich zugängliche Quellen nachprüfen.

Bei einigen Online-Portalen herrscht hingegen eine große Volatilität, so dass die von uns verwendeten Ausgangsdaten nicht mehr unbedingt verfügbar sind. Belege dafür, dass die von uns verwendeten Ausgangsdaten tatsächlich die korrekten sind, können Sie in diesen Fällen aber direkt von uns über die in Fußnote 1 angegebene Korrespondenzadresse erhalten.

Und die Überprüfung der mathematischen Richtigkeit unserer Schlussfolgerungen durch Sie, die Leserinnen und Leser, erfordert nichts weiter als ein bisschen Hauptschulmathematik.

2 Worum geht es?

Millionen Kaufentscheidungen werden jeden Tag aufgrund von rankingbasierten Tests in Zeitungen, Zeitschriften und anderen Medien getroffen. Menschen entscheiden nach Rankings, wo sie leben, arbeiten oder studieren werden. Schwerwiegende politische und ökonomische Entscheidungen werden durch nationale und internationale Rankings massiv beeinflusst. Und so weiter.

In dieser Arbeit zeigen wir, dass Rankings inhärent problematisch sind, auch wenn alle Beteiligten guten Willens sind und nach besten Standards vorgehen.

Unsere Arbeit ist nicht theoretisch-akademisch, sondern wir gehen mitten ins Leben. Konkret haben wir Rankings aus sieben beispielhaften Medien mit hoher Reichweite ausgewählt:

- *Stiftung Warentest* (www.test.de)
- *CHIP Online* (www.chip.de): wird regelmäßig zitiert u.a. in computerbild.de
- *fotoMAGAZIN* (www.fotomagazin.de)
- *ETM TESTMAGAZIN* (etm-testmagazin.de/tests)
- *Haus & Garten Test* (www.haus-garten-test.de)
- *Shanghai-Ranking Universitäten* (www.shanghairanking.com/ARWU2012.html)
- *CHE-Hochschulranking* (ranking.zeit.de/che2012/de/)

Abgesehen vom CHE-Hochschulranking basieren alle diese Tests darauf, dass mehrere Kriterien mit unterschiedlichen Gewichtungen in einem Gesamtergebnis zusammengerechnet werden. Wir müssen feststellen, dass schon kleine Änderungen an den in einem Test verwendeten Gewichtungen zu durchaus nicht vernachlässigbaren Änderungen am Ergebnis führen. Konkret gibt es immer wieder ernsthafte Änderungen an den Rankingplätzen einzelner Testobjekte auch schon dann, wenn jedes einzelne Gewicht um nicht mehr als *maximal* einen Prozentpunkt vermindert oder um *maximal* zwei Prozentpunkte erhöht wird. Dies ist bedeutsam, da die Festlegung der Gewichtungen – auf ein bis zwei Prozentpunkte genau – nicht sachlich zu begründen, sondern bis zu einem gewissen Grad immer willkürlich ist. Was man allein schon daran sieht, dass die Gewichtungen in der Regel ganze Vielfache von fünf oder zehn Prozentpunkten sind – nicht aufgrund von sachgerechten Überlegungen, sondern schlicht weil das Dezimalsystem die Basis ist.

Das CHE-Hochschulranking ist ein Sonderfall. Es basiert auf einer komplexeren Methode, die diese Problematik zu umgehen sucht. Allerdings mussten die Entwickler auch dieses Rankings mindestens eine willkürliche Entscheidung treffen, um am Ende zu einer Reihung zu gelangen. Und auch hier zeigt sich wieder, dass einzelne wenige Hochschulen an ganz anderer Stelle in der Wertung landen würden, wenn diese Entscheidung etwas anders – aber ebenso plausibel – getroffen worden wäre. Aufgrund dieser grundsätzlich anderen Vorgehensweise werden wir das CHE-Hochschulranking nicht gemeinsam mit den anderen Medien, sondern separat in Abschnitt 10 behandeln.

Nicht untersucht haben wir Ökotest (www.oekotest.de). Das hat einen positiven und einen negativen Grund. Positiv ist, dass die Methodik von Ökotest oft sehr viel zurückhaltender ist und nicht die typischen Angriffsflächen anderer Testverfahren bietet. Negativ hingegen ist, dass doch noch potentielle Schwachstellen bestehen könnten, aber mangels Offenlegung der Zwischenergebnisse nicht von uns analysiert werden konnten. Siehe Abschnitt 11 für genauere Erläuterungen zu Ökotest.

3 Wie funktionieren Rankings?

Abgesehen vom CHE-Hochschulranking gehen alle untersuchten Rankingmethoden nach demselben Grundprinzip vor und sind damit stellvertretend für unzählige weitere Rankings. Zur Veranschaulichung beginnen wir mit zwei konkreten Beispielen.

Beispiele:

- Beim Test von Tablets bei CHIP Online wird das Kriterium *Handling* mit 50%, das Kriterium *Mobilität* mit 20%, das Kriterium *Display* ebenfalls mit 20% und das Kriterium *Ausstattung* mit 10% gewichtet. Das fiktive Produkt „Tablet XYZ“ möge beim ersten Kriterium 88 Punkte erzielen, beim zweiten 70, beim dritten 90 und beim vierten 75 Punkte. Als Gesamtergebnis für Produkt „Tablet XYZ“ ergeben sich
$$0{,}5 \cdot 88 + 0{,}2 \cdot 70 + 0{,}2 \cdot 90 + 0{,}1 \cdot 75 = 83{,}5 \text{ Punkte.}$$
- Stiftung Warentest vergibt Schulnoten statt Punktzahlen. Beim Test von Bodenstaubsaugern in der Ausgabe 4/2012 (Seite 62) wird die Note für *Saugen* mit 45%, die Note für die *Handhabung* mit 30%, die Note *Umwelteigenschaften* mit 15% und die Note *Haltbarkeit* mit 10% gewichtet. Für das fiktive Produkt „Bodenstaubsauger XYZ“ sei die erste Note eine 1,8 die zweite eine 2,0, die dritte eine 4,6 und die vierte eine 1,4. Als Gesamtnote für Produkt „Bodenstaubsauger XYZ“ ergibt sich ohne Sonderregeln (s.u.) zunächst einmal
$$0{,}45 \cdot 1{,}8 + 0{,}3 \cdot 2{,}0 + 0{,}15 \cdot 4{,}6 + 0{,}1 \cdot 1{,}4 = \text{Note } 2{,}2.$$

Bei den allermeisten Rankings kommt das Endergebnis ähnlich wie in diesen beiden Beispielen zustande: Jedes untersuchte Testobjekt wird nach mehreren Kriterien bewertet, und jedes Einzelkriterium ergibt eine Einzelpunktzahl oder Einzelnote für das Testobjekt. Jedes Kriterium erhält ein individuelles Gewicht, meist angegeben in Prozentpunkten, so dass sich alle Gewichte zu 100% aufsummieren. Um die Gesamtpunktzahl bzw. Gesamtnote für ein Testobjekt zu berechnen, wird jede Einzelpunktzahl bzw. Einzelnote mit dem Gewicht des Kriteriums multipliziert, und alle diese Teilergebnisse werden aufaddiert. Nach dem Wert dieser Summe werden die Testobjekte schlussendlich im Ranking gereiht.

In manchen Medien wie bspw. bei Stiftung Warentest kommen noch Sonderregeln bei der Berechnung des Endergebnisses aus den Einzelergebnissen zum Einsatz. Zur Illustration setzen wir das zweite Beispiel von eben fort.

Fortsetzung zweites Beispiel:

- Wie oben gesagt, würde das fiktive Produkt „Bodenstaubsauger XYZ“ eigentlich die Gesamtnote 2,2 erhalten. Aber in diesem Testbericht findet sich in der Fußnote „Führt zur Abwertung“ auf Seite 64 die Sonderregel,[2] dass die berechnete Gesamtnote um eine ganze Note herabgesetzt wird, falls die *Umwelteigenschaften* nur *mangelhaft* sind. Da „Bodenstaubsauger XYZ“ für seine Umwelteigenschaften im obigen Beispiel tatsächlich nur ein *Mangelhaft* (4,6) erhalten hat, ergibt sich als endgültige Gesamtnote also nur noch eine 3,2.

4 Was genau ist das Problem?

4.1 Generelle Problematik von Rankings

Es geht uns um diese Gewichte, also etwa in in Abschnitt 3 um die 50%, 20%, 20% und 10% im Beispiel aus CHIP Online bzw. um die 45%, 30%, 15% und 10% im Beispiel aus Stiftung Warentest. Selbstverständlich gehen wir davon aus, dass die Autoren jeder rankingbasierten Studie diese Gewichte sehr sorgfältig und sachorientiert festlegen. Dass dennoch eine gewisse Willkür unvermeidlich ist, zeigt sich schon in den „runden“ Werten 10%, 20%, 25% 30% usw.: Dass beispielsweise genau 20% und nicht 19% oder 22% festgelegt sind, ist offenkundig nicht das Ergebnis fundierter sachorientierter Erwägungen, sondern einfach dem Dezimalsystem geschuldet.

Da der Unterschied beispielsweise zwischen 19%, 20% und 22% nicht sachlich zu begründen ist, wäre die Aussagekraft von Rankings im Detail infrage zu stellen, falls sich herausstellen würde, dass schon der Unterschied zwischen 19%, 20% oder 22% zu nicht vernachlässigbaren Unterschieden bei den Ergebnissen führt.

Wir haben untersucht, was sich ändert, wenn jedes Gewicht um maximal einen Prozentpunkt verkleinert oder um maximal zwei Prozentpunkte vergrößert wird, und sind zu einem durchaus problematischen Ergebnis gekommen.

Fortsetzung erstes Beispiel aus Abschnitt 3:

- Die Gewichte bei diesem beispielhaft angeführten Test waren 50%, 20%, 20% und 10%. Wir untersuchen also Variationen, bei denen das Gewicht für das erste Kriterium zwischen 49% und 52% rangiert, die Gewichte für das zweite und dritte Kriterium unabhängig voneinander zwischen 19% und 22% und das Gewicht für das vierte Kriterium zwischen 9% und 12%. Natürlich müssen alle so variierten Gewichte sich wieder zu 100% aufsummieren. Eine mögliche Variation der Gewichte, die all dies erfüllt, wäre etwa 49%, 22%, 19% und 10% (dies ist nur ein illustratives Beispiel und nicht die in unserer Studie verwendete Variation).

[2] Besagte Fußnote enthält noch eine zweite Sonderregel, die in unserem Beispiel aber nicht greift und daher hier außen vor bleiben kann.

4.2 Spezielle Problematik von Schulnoten

Wie oben gesagt, vergeben einzelne Medien am Ende Schulnoten; unter den von uns untersuchten Medien tun das nur Stiftung Warentest und Haus und Garten. Die Noten werden hier auf Zehntel genau errechnet und dann auf die aus der Schule bekannte grobe Skala *sehr gut*, *gut*, *befriedigend*, *ausreichend* und mangelhaft umgerechnet. Die Note sehr gut wird bei Stiftung Warentest bis zur Notenstufe 1.5, die Note gut von 1.6 bis 2.5, befriedigend von 2.6 bis 3.5 und ausreichend von 3.6 bis 4.5 vergeben, alles ab 4.6 ist mangelhaft. Bei Haus und Garten wechselt die Note hingegen zwischen 1.4 und 1.5, zwischen 2.4 und 2.5, zwischen 3.4 und 3.5 sowie zwischen 4.4 und 4.5.

In einem solchen Fall ist natürlich diese Schulnote das Hauptergebnis, nicht die genaue Reihenfolge der Produkte, wie wir sie hier untersuchen. Tatsächlich haben wir nur relativ wenige Sprünge von einer dieser Schulnoten zu einer anderen gefunden. In allen von uns betrachteten Tests zusammengenommen, haben sich sowohl bei Stiftung Warentest als auch bei Haus und Garten jeweils nur bei drei Produkten überhaupt verschiedene Noten zwischen dem ursprünglichen Test und unserer Variation der Gewichtungen ergeben. Darunter ist allerdings in einem der Einzeltests von Stiftung Warentest und in zwei Einzeltests von Haus und Garten jeweils ein herausragender Fall: Das Spitzenprodukt, das als einziges mit sehr gut bewertet wurde, wechselt mit einer unserer variierten Gewichtungen auf gut, so dass kein einziges Produkt in diesem Test mehr ein sehr gut erhält.

Bei welchen Notenstufen man die Grenzen zwischen den Noten zieht, liegt im Prinzip im Ermessen des Notengebers. Es gibt da keine „absolute Wahrheit“, was sich ja auch schon darin zeigt, dass die Grenzen bei *Stiftung Warentest* und bei *Haus und Garten* nicht identisch sind. Daher haben wir zusätzlich betrachtet, was passieren würde, wenn die Grenzen statt dessen bei den exakten Werten 2,0, 3,0, 4,0 und 5,0 gezogen werden, also alles bis 1,9 ist *sehr gut*, alles von 2,0 bis 2,9 ist *gut*, alles von 3,0 bis 3,9 ist *befriedigend*, alles von 4,0 bis 4,9 ist *ausreichend* und alles weitere ist *mangelhaft*. Das mag vielleicht etwas ungewohnt sein, wäre aber ebenfalls plausibel. Wenn man so rechnet, ergeben sich bei Stiftung Warentest und bei Haus und Garten jeweils zwei Notenwechsel über alle Tests hinweg. Darunter ist bei Stiftung Warentest allerdings ein herausragender Fall: Zwei Testsieger sind im Original beide gut, mit einer unserer variierten Gewichtungen wechselt einer von beiden zu sehr gut und wird damit (1) alleiniger Testsieger und (2) einziges Produkt mit Note sehr gut.

5 Was genau haben wir untersucht?

Folgende rankingbasierten Tests haben wir uns beispielhaft vorgenommen:

- *Stiftung Warentest*: alle neun Tests mit mindestens zehn Produkten im Zeitraum März-Juni 2012 sowie zum Vergleich ein Test mit nur sechs Produkten.
- *CHIP Online*: drei zufällig ausgewählte Tests, die am Stichtag 29. Oktober 2012 auf dem Portal mit mindestens fünfzehn und höchstens vierzig Produkten publiziert waren, sowie ein weiterer Test mit nur acht Punkten zum Vergleich.
- *fotoMAGAZIN*: zwei zufällig ausgewählte Tests auf der Webseite am Stichtag 29. Oktober 2012.
- *ETM TESTMAGAZIN*: alle Tests mit mehr als einem Produkt aus der Ausgabe 07/2012.
- *Haus & Garten Test*: drei zufällige Tests mit mindestens fünf Produkten aus der Ausgabe 04/2012.
- *Shanghai*: die 21 höchstplatzierten Universitäten im Ranking 2012.

- *CHE*: Hochschulranking 2012/13, daraus ausgewählt die universitären Disziplinen Anglistik, Biologie, Chemie, Geographie, Kommunikationswissenschaft, Wirtschaftsingenieurwesen und Zahnmedizin.

Alle diese Medien wurden ausgewählt, weil sie einflussreiche, hoch angesehene Primärquellen für rankingbasierte Studien sind. Die ausgewählten Ausgaben der Zeitschriften bzw. Stichtage bei Online-Portalen waren einfach die aktuellen zum Zeitpunkt der Bearbeitung durch uns. Die einzelnen Tests wurden ohne vorherige Ansicht ausgewählt, um eine tendenziöse Zusammenstellung zu vermeiden.

Wir wollen in diesem Papier keine Behauptungen beweisen, sondern im Gegenteil „nur" Gewissheiten erschüttern. Für diese – sehr viel bescheidenere – Zielsetzung sind die untersuchten Tests unseres Erachtens ausreichend zufällig ausgewählt und daher als ausreichend breit und repräsentativ einzustufen.

Im Einzelnen haben wir die Tests aus der folgenden Tabelle untersucht. Bei Printmedien ist jeweils Ausgabe und Seitenzahl aufgeführt, bei Online-Medien noch einmal das Datum unseres Zugriffs.

Stiftung Warentest	Rote Nagellacke	6/2012, S. 32
	Smartphones	5/2012, S. 42
	Wandfarben	5/2012, S. 56
	City-Fahrradhelme	4/2012, S. 72
	Farbschutzschampoos	4/2012, S. 34
	Drucker-Scanner	4/2012, S. 54
	Bodenstaubsauger	4/2012, S. 62
	Elektrovertikulierer	3/2012, S. 72
	Multivitaminsaft	3/2012, S. 24
	Sommerreifen	3/2012, S. 74
CHIP Online	3D-Fernseher 43 Zoll	29.10 2012
	MP3-Player	
	Notebooks 14-15,6 Zoll	
	Tablets	
ETM Testmagazin	Kochtöpfe	7/2012, S. 26
	Standmixer	7/2012, S. 66
	Dörrgeräte	7/2012, S. 94
Haus & Garten Test	Vakumierer	4/2012, S. 32
	Kohlegrills	4/2012, S. 40
	Gasgrills	4/2012, S. 48
fotoMagazin	Kompaktkameras	29.10.2012
	Superzoomkameras	
Shanghai-Ranking	Ranking 2012	–––
CHE-Ranking	Ranking 2012	–––

6 Detaillierte Daten

Im Zusatzpapier „Sind Rankings inhärent willkürlich? – Detaillierte Daten“ finden Sie die Daten, aus denen die Ergebnistabellen in Abschnitt 8 und 9 erzeugt wurden:

- In Abschnitt 1 zu jedem Test das Originalranking im unmittelbaren Vergleich mit unserem Ranking auf Basis unserer variierten Gewichtungen.
- In Abschnitt 2.1 und 2.2 wesentliche Informationen zum Zustandekommen der Daten.
- In Abschnitt 2.3 und 2.4 dann die variierten Gewichtungen selbst.

7 Wie sind wir vorgegangen?

In Abschnitt 7-9 klammern wir das CHE-Hochschulranking weiterhin aus.

7.1 Unsere generelle Vorgehensweise

Eine systematische Auflistung unserer Ergebnisse gemäß der im Folgenden beschriebenen generellen Vorgehensweise finden Sie in Abschnitt 9.1.

In jedem der untersuchten Tests gibt es eine Reihe von Kriterien, und jedes Kriterium hat ein Gewicht in Prozentpunkten, die sich zu 100 aufsummieren. Diese Gewichte haben wir hergenommen und durch ein selbst entwickeltes kleines Computerprogramm eine Million Mal auf unterschiedliche Weise zufällig variiert. Dabei haben wir darauf geachtet, dass kein Gewicht um mehr als einen Prozentpunkt vermindert oder um mehr als zwei Prozentpunkte erhöht wird und natürlich in Summe wieder 100% herauskommt. Bei jeder dieser eine Million Variationen der Gewichtungen haben wir geprüft, was mit den Rankingergebnissen passiert, und haben dann für jeden Test die uns am interessantesten erscheinende Variation hergenommen. Unser Kriterium dafür, wie interessant eine Variation ist, ist der *Durchschnittswert*, um wie viele Plätze jedes Produkt seine Position im Ranking durch diese Variation geändert hat.

7.2 „Echte“ Veränderungen

Wenn zwei Produkte dieselbe Gesamtpunktzahl bzw. Gesamtnote erreichen, weisen wir ihnen denselben Platz im Ranking zu. Auf Platz 1 sind daher alle Produkte, die die beste überhaupt erzielte Gesamtpunktzahl bzw. Gesamtnote erreicht haben, auf Platz 2 alle Produkte mit zweitbester Gesamtpunktzahl bzw. Gesamtnote usw.

Zur Illustration nehmen wir ein Beispiel vorweg: der Smartphone-Test von Stiftung Warentest. In der zweiten Spalte sind die Produkte gemäß der realen Testergebnisse gereiht, in der dritten Spalte gemäß einer unserer variierten Gewichtungen:

Platz	Originale Gewichtung	Variierte Gewichtung
1	Motorola Razar (xt910) Samsung Galaxy Note	Samsung Galaxy Note
2	HTC Rhyme Samsung Galaxy Nexus Sony Ericsson Xperia arc S	Motorola Razar (xt910)

	Sony Ericsson Xperia Neo V	
3	Blackberry Torch 9860 HTC Sensation XE HTC Sensation XL	Samsung Galaxy Nexus Sony Ericsson Xperia arc S Sony Ericsson Xperia Neo V
4	HTC Radar HTC Titan Samsung Wave 3 Sony Ericsson Xperia ray	HTC Rhyme
5	HTC Explorer Motorola Defy+ Nokia 500	Blackberry Torch 9860 HTC Sensation XL Samsung Wave 3
6	Blackberry Curve 9360 Nokia 700 Nokia Lumia 710 Nokia Lumia 800	HTC Radar HTC Sensation XE HTC Titan Sony Ericsson Xperia ray
7	Samsung Galaxy Y	Motorola Defy+
8		Blackberry Curve 9360 HTC Explorer Nokia 500 Nokia 700 Nokia Lumia 800
9		Nokia Lumia 710
10		Samsung Galaxy Y

Man sieht mit bloßem Auge, dass Produkte „springen". Dadurch, dass mehrere Produkte auf demselben Platz sein können, ist die Situation allerdings etwas unübersichtlich. Es macht wenig Sinn, einfach die Plätze, auf denen ein Produkt in den beiden Rankings landet, zu vergleichen. Denn dann würde beispielsweise für Samsung Galaxy Y ein Sprung um drei Plätze von 7 nach 10 herauskommen, obwohl dieses Produkt in beiden Rankings das einsame Schlusslicht ist. Einfacher Vergleich der Plätze ist offenbar kein gutes Vergleichsmaß.

Um diese Sprünge und die *Sprungweiten* in angemessener Weise zu messen, vergleichen wir daher nicht die Plätze eines Produkts in den beiden Rankings, sondern das, was wir *bereinigte Plätze* nennen:

Der *bereinigte Platz* eines Produkts in einem Ranking zählt die Anzahl der Produkte auf den Plätzen vor diesem Produkt plus 1.

Im Fall, dass alle Produkte unterschiedliche Plätze einnehmen, ist der so definierte *bereinigte Platz* identisch mit dem, was man üblicherweise unter „Platz" versteht.

Illustratives Beispiel aus der obigen Tabelle:

- Das Produkt HTC Rhyme ist im ursprünglichen Ranking auf Platz 2 und hat daher zwei Produkte vor sich, die sich Platz 1 teilen. Somit ist HTC Rhyme auf dem *bereinigten* Platz 2+1=3. Im Ranking mit unserer variierten Gewichtung ist HTC Rhyme auf Platz 4, das heißt, die insgesamt fünf Produkte auf den Plätzen 1-3 sind vor HTC Rhyme. Daher hat HTC Rhyme den bereinigten Platz 5+1=6.

Die *Sprungweite* eines Produkts beim Vergleich zweier Rankings ist die Differenz der beiden bereinigten Plätze (in den Betrag genommen).

Beispiel fortgesetzt:

- HTC Rhyme hat also Sprungweite |3-6|=3. Der Betrag wird genommen, um Sprünge nach oben und Sprünge nach unten gleich zu zählen.

Diese Sprungweiten von Produkten sind natürlich das interessanteste Ergebnis und werden dementsprechend viel Aufmerksamkeit bei der Präsentation der Ergebnisse in Abschnitt 8 und 9.1 finden.

Daneben sticht in der Tabelle oben ein weiterer Aspekt ins Auge: Insgesamt 21 Produkte verteilen sich auf nur sieben verschiedene Plätze im originalen Ranking. Das heißt, das Ranking ist nicht sehr trennscharf. Dies trifft auf etliche von uns untersuchte Rankings zu. Dass wir dennoch in Abschnitt 8 und 9.1 auch bei diesen Rankings von einer recht großen Anzahl von Sprüngen berichten müssen, ist umso bemerkenswerter.

Hinzu kommt aber ein weiterer dokumentierungswürdiger Effekt: Manche Produkte sind im originalen Ranking auf demselben Platz, in unserem Ranking aber auf verschiedenen Plätzen oder umgekehrt. Beispielsweise teilen sich Motorola Razar (xt910) und Samsung Galaxy Note im originalen Ranking in der Tabelle oben den ersten Platz, im variierten Ranking hingegen haben sie unterschiedliche Plätze nämlich Platz 1 und 2.

Wir werden in Abschnitt 9.1 daher zusätzlich bei jedem Produkttest zählen, wie viele Produkte im originalen Ranking auf demselben Platz und in unserem variierten Ranking auf unterschiedlichen Plätzen landen bzw. wie viele Produkte im originalen Ranking auf unterschiedlichen Plätzen und im variierten Ranking auf demselben Platz landen.

7.3 „Alphabetische“ Veränderungen

Wenn ein Rankingergebnis in einer Tabelle präsentiert wird, werden Produkte, die auf demselben Platz gelandet sind, in manchen Medien in alphabetischer Reihenfolge nacheinander aufgelistet. Da sehr häufig sehr viele Produkte denselben Platz einnehmen, kann ein Produkt an sehr unterschiedlichen Positionen platziert sein, je nachdem, ob sein Name mit A oder mit Z beginnt. Das ist formal nicht zu beanstanden, die beistehenden Erläuterungen sowie die angegebenen Punktzahlen bzw. Schulnoten reichen in der Regel aus, um die Reihenfolge der Produkte korrekt zu interpretieren.

Allerdings ist allgemein bekannt, dass Menschen sich – auch wider besseren Wissens – stark von optischen Eindrücken leiten lassen und daher die genaue Platzierung eines Produktes in einer Auflistung allein durch die Optik schon einen Einfluss haben kann. Aus diesem Grund haben wir zusätzlich bei jedem betroffenen Test untersucht, wie weit die Produkte bei der gängigen Art der alphabetischen Auflistung ihre Position verändern, wenn wir die Gewichte nach unserer in Abschnitt 7.1 vorgestellten Methode variieren.

Die Analyse der alphabetischen Veränderungen finden Sie in Abschnitt 9.2. Unter den von uns untersuchten Medien betrifft dieser Aspekt nur Stiftung Warentest, CHIP Online und fotoMagazin, die anderen Medien reihen nicht alphabetisch.

Da bei alphabetischer Einordnung ranggleicher Produkte definitionsgemäß keine zwei Produkte auf demselben Platz landen können, brauchen Plätze nicht bereinigt zu werden, und wir nehmen in Abschnitt 9.2 als Sprungweite einfach die Differenz der beiden Plätze (wieder in den Betrag genommen).

Durch die alphabetische Einordnung unter Punkt-/Notengleichen kann der Sprung eines Produkts im Einzelfall größer oder kleiner sein als der Sprung allein durch Punkte-/Notengleichheit, den wir in unserer oben beschriebenen generellen Vorgehensweise betrachten.

7.4 Rundungsprobleme und andere Diskrepanzen zu den Originalergebnissen

Wir sind den Erläuterungen der Berechnungsmethoden in der jeweiligen Publikation penibel gefolgt. Dennoch haben wir nicht bei jedem Produkt in jedem Ranking exakt das publizierte Ergebnis nachvollziehen können.

- Für den Test von Smartphones durch *Stiftung Warentest* sind die Abwertungsregeln nicht exakt angegeben.
- Bei *CHIP Online* konnten wir generell die angewendeten (aber nicht publizierten) Rundungsregeln nicht exakt nachbilden, obwohl wir diverse plausible Rundungstechniken ausprobiert hatten. Vereinzelte kleine Diskrepanzen in anderen Medien scheinen ebenfalls auf unterschiedliche Rundung zurückzuführen sein.
- Bei mindestens einem Produkt scheint die publizierte Punktzahl durch einen Tippfehler verfälscht zu sein („B/R/K cookvision alpha Kochtopf 3l" beim Test von Kochtöpfen im *ETM TESTMAGAZIN*).

In Abschnitt 2.1 des Zusatzpapiers „Sind Rankings inhärent willkürlich? – Detaillierte Daten" sind alle Produkte in allen Tests aufgelistet, bei denen sich Diskrepanzen gezeigt haben.

Selbstverständlich haben wir in allen diesen Fällen nicht die publizierten Testergebnisse selbst, sondern unsere leicht davon abweichenden Ergebnisse zum Vergleich mit den von uns selbst generierten Variationen hergenommen (bei den nichttransparenten Abwertungsregeln heißt das, dass wir gar nicht abgewertet haben).

Dieses Vorgehen ist absolut gerechtfertigt, denn wichtig für unsere Untersuchung ist nicht, welche Rundungsregel wir wählen, sondern dass Endergebnisse, die wir miteinander vergleichen, auf genau dieselbe Art gerundet wurden.

Ein Kommentar speziell zum Shanghai-Ranking: Die publizierten Daten sind auf 100 skaliert, das heißt, die beste Universität hat immer 100 Punkte, auch wenn sie nicht volle Punktzahl erreicht hat. In Abschnitt 1.1 und 1.2 von „Sind Rankings inhärent willkürlich? – Detaillierte Daten" präsentieren wir die unskalierten Rohergebnisse. Und in Abschnitt 2.1 ebendort belegen wir die Übereinstimmung mit den publizierten skalierten Ergebnissen.

Die von uns gefundenen krassesten Fälle sind ohnehin nicht betroffen, und unsere statistischen Ergebnisse sind auch nur marginal beeinflusst. Die Kernaussage dieses Papiers ist also durch diese Diskrepanzen nicht infrage zu stellen.

8 Und was kam heraus – die markantesten Einzelfälle für den schnellen Leser

Wir betrachten hier besonders markante Beispielfälle aus Stiftung Warentest, CHIP Online, ETM TESTMAGAZIN, Haus & Garten Test sowie fotoMagazin.

8.1 Sprünge in den Reihungen

Bei einigen der ausgewählten Tests sind die Sprünge *ohne* alphabetische Sortierung markant, bei anderen die Sprünge *mit* alphabetischer Sortierung, bei vier Tests sogar beide. Zuerst die beiden Tabellen, Erläuterung und Auswertung folgen danach.

Ausgewählte markante Fälle *ohne* alphabetische Effekte (Vorgehen gemäß Abschnitt 7.2):

Konkreter Test	**Anzahl getestete Produkte**	**Summe aller Sprungweiten *ohne* alphabetische Effekte**	**Durchschnittliche Sprungweite**	**Sprungweiten im Einzelnen**
Stiftung Warentest: Multivitaminsaft	20	5	0.25	1x5
Stiftung Warentest: City-Fahrradhelme	16	6	0.37	6x1
CHIP Online: 3D-Fernseher 43''	28	33	1.17	2x5, 1x4, 1x3, 4x2, 8x1
CHIP Online: Notebooks 14-15.6''	17	15	0.88	1x3, 3x2, 6x1
ETM TESTMAGAZIN: Kochtöpfe	24	17	0.7	2x3, 2x2, 7x1
Haus & Garten Test : Vakuumierer	6	3	0.5	1x2, 1x1
fotoMagazin: Superzoomkameras	16	16	1.0	2x4, 1x2, 6x1

Ausgewählte markante Fälle *mit* alphabetischen Effekten (Vorgehen gemäß Abschnitt 7.3):

Wie schon in Abschnitt 7.3 erwähnt, betrifft dieser Punkt nur Stiftung Warentest, CHIP Online und fotoMagazin, so dass alle Beispiele aus diesen drei Medien entnommen sind.

Konkreter Test	**Anzahl getestete Produkte**	**Summe aller Sprungweiten *mit* alphabetischen Effekten**	**Durchschnittliche Sprungweite**	**Sprungweiten im Einzelnen**
Stiftung Warentest: Smartphones	21	26	1.23	3x3, 3x2, 11x1
Stiftung Warentest: Multivitaminsaft	20	10	0.50	1x5, 5x1
Stiftung Warentest: City-Fahrradhelme	16	8	0.50	1x3, 5x1
CHIP Online: 3D-Fernseher 43''	28	32	1.14	2x5,1x4,1x3,3x2, 9x1
CHIP Online: Notebooks 14-15.6''	17	14	0.82	1x3, 2x2, 7x1
CHIP Online: Tablets	40	28	0.7	1x3,6x2, 13x1
fotoMAGAZIN: Superzoomkameras	16	16	1.00	2x3, 4x2, 2x1

Erläuterung und Auswertung:

Zunächst zur Erzeugung der Ergebnisse: Die Variation der Gewichtungen, mit denen die Ergebnisse in den beiden obigen Tabellen erzeugt wurden, sind für jeden Test identisch mit der Variation der Gewichtungen bei der Analyse des Gesamttests in Abschnitt 9.1.2 bzw. 9.2.2. Detailerläuterungen dazu finden Sie auch in diesen beiden Abschnitten. Die Einzeltabellen finden Sie in Abschnitt 1.2 bzw. Abschnitt 1.4 von „Sind Rankings inhärent willkürlich? – Detaillierte Daten" und die verwendeten variierten Gewichtungen in Abschnitt 2.3.2 und 2.3.4 ebendort.

Nun zu den Tabellen selbst: Neben der Identifikation des jeweiligen Tests und der Anzahl der getesteten Produkte finden Sie eine Spalte „Summe aller Sprungweiten *ohne* alphabetische Effekte / *mit* alphabetischen Effekten". Für jedes Produkt, das seine Position ändert, messen wir, um wie viele Positionen es springt. Die Summe aller Sprungweiten aller Produkte wird in dieser Spalte aufsummiert. In der letzten Spalte werden dieselben Sprungweiten dann noch einmal im Einzelnen aufgelistet.

Beispiel:

- Beim Test von Multivitaminsaft ohne alphabetische Effekte (erste Tabelle oben, erste Zeile) ist genau ein Produkt um fünf Positionen gesprungen, eines um vier, und sechs Produkte sind jeweils um eine Position gesprungen. Als Summe aller Sprungweiten ergibt sich $1 \cdot 5 + 1 \cdot 4 + 6 \cdot 1 = 15$, daher steht in der dritten Spalte eine 15.

Die vorletzte Spalte ist das Kriterium, nach dem wir diese besonders markanten Fälle ausgewählt haben: der Quotient aus der Summe der Sprungweiten (dritte Spalte), dividiert durch die Anzahl der getesteten Produkte (zweite Spalte).

Markanteste Beispiele:

- In der Tabelle *ohne* alphabetische Effekte tritt die höchste Zahl in der vorletzten Spalte beim Test von 3D-Fernsehern durch CHIP Online auf: 1.17. Der zweithöchste Wert (Superzoomkameras bei fotoMagazin) ist immerhin noch genau 1.0.
- In der Tabelle *mit* alphabetischen Effekten kommen zwei Tests sogar auf noch höhere Werte: 1.24 bei Smartphones (Stiftung Warentest) und 1.14 wieder bei 3D-Fernsehern. Auch hier ist der Wert für Superzoomkameras genau 1.0.

Machen Sie sich klar, was diese Zahlen konkret bedeuten:

Bei zweien der von uns zufällig ausgewählten Tests springt jedes Produkt sowohl *ohne* als auch *mit* alphabetischen Effekten im Durchschnitt um eine Position oder sogar um mehr als eine Position. Bei einem weiteren Test springt jedes Produkt nur *mit* alphabetischen Effekten im Durchschnitt um mehr als eine Position.

Beachten Sie, dass bei der Betrachtung *ohne* alphabetische Effekte insgesamt nur 22 – nicht tendenziös ausgewählte – Tests überhaupt einbezogen waren, bei der Betrachtung *mit* alphabetischen Effekten sogar nur 16 Tests. Davon sind insgesamt drei sicherlich keine vernachlässige Größe. Zudem sind die durchschnittlichen Sprungweiten bei den anderen Tests in den beiden Tabellen – 0.37...0.75 in der ersten und 0.5...0.82 in der zweiten – ebenfalls ganz und gar nicht unproblematisch, wenn auch nicht ganz so herausragend. Bemerkenswert sind zudem die einzelnen Sprungweiten in der letzten Spalte, die wir als nächstes genauer betrachten.

Markanteste Beispiele:

- *Ohne* alphabetische Effekte: Bei Multivitaminsäften etwa springt ein Produkt gleich um fünf Positionen, ein zweites um vier. Bei 3D-Fernsehern springen zwei Produkte um fünf Positionen, ein weiteres um vier und noch eins um drei.
- *Mit* alphabetischen Effekten: Beim Test von Multivitaminsäften springt ein Produkt gleich um sechs Positionen, beim Test von 3D-Fernsehern springen zwei Produkte um jeweils fünf Positionen, ein weiteres um vier Positionen.

8.2 Notensprünge

Vergleiche Abschnitt 4.2. Wie schon erwähnt, bilden unter den von uns untersuchten Medien nur *Stiftung Warentest* und *Haus und Garten Test* Schulnoten. Wir haben gezielt für die Erzeugung von Notensprüngen weitere variierte Gewichtungen berechnet. Mit den in *Stiftung Warentest* und *Haus und Garten Test* verwendeten Notenschemata konnten wir folgende Sprünge erzeugen:

1. *Stiftung Warentest* / Smartphones: Motorola Defy+ von *befriedigend* nach *gut*.
2. *Stiftung Warentest* / City-Fahrradhelme: Nutcase URS-011S von *gut* nach *befriedigend*.
3. *Stiftung Warentest* / Multivitaminsaft: Rabenhorst 11 plus 11: von *sehr gut* nach *gut*.
4. *Haus und Garten* / Vakumierer: Gastroback Design Pro Vakumierer von *sehr gut* nach *gut*.
5. *Haus und Garten* / Kohlegrills: Barbecook Major Inox von *sehr gut* nach *gut*.
6. *Haus und Garten* / Gasgrills: Campingaz Genesco Classis 3L: von *sehr gut* nach *gut*.

Der dritte, fünfte und sechste Fall sind herausragend. In allen drei Fällen ist das abgewertete Produkt sowohl mit den originalen als auch mit unseren variierten Gewichtungen der alleinige Testsieger. Mit den variierten Gewichtungen gibt es überhaupt kein *sehr gut* mehr, auch der Testsieger ist nur noch *gut*.

Wie ebenfalls schon in Abschnitt 4.2 angesprochen, haben wir auch den Fall betrachtet, dass die Grenzen zwischen den Schulnoten anders, aber ebenfalls plausibel gezogen sind. Konkret haben wir dafür die Grenze zwischen *sehr gut* und *gut* nicht zwischen 1,5 und 2,5, sondern zwischen 1,9 und 2,0 gesetzt, und die anderen Grenzen entsprechend. Das heißt, alles ab 2,0 ist *gut*, alles ab 3,0 *befriedigend* usw. Mit diesem Notenschema würden sich folgende Sprünge ergeben:

1. *Stiftung Warentest* / Smartphones: Samsung Galaxy Note von *gut* nach *sehr gut*.
2. *Stiftung Warentest* / Wandfarben: Düfa Superweiss plus von *sehr gut* nach *gut*.
3. *Haus und Garten* / Kohlegrills: Landmann Maximo 31261 von *sehr gut* nach *gut*.
4. *Haus und Garten* / Gasgrills: Weber Q 220 Black Line Station von *sehr gut* nach *gut*.

Der erste Fall ist herausragend: Mit der originalen Gewichtung gibt es zwei punktgleiche Testsieger, Samsung Galaxy Note und Motorola Razar (xt910), beide *gut*. Mit unserer variierten Gewichtung setzt Samsung Galaxy Note sich ab, wird alleiniger Testsieger und ist als einziges Produkt *sehr gut*.

Allein die in diesem Abschnitt präsentierten Fallbeispiele reichen schon aus, um die Kernthese dieses Papiers zu belegen: dass eine nicht vernachlässigbare Anzahl von Einzeltests ein so problematisches Ergebnis zeigen, dass die Aussagekraft von Rankings durchaus infrage steht.

Im nun folgenden Abschnitt 9 zeigen wir, dass auch die anderen Einzeltests in der Mehrzahl nicht unproblematisch sind, auch wenn sie nicht ganz so markant herausstechen. Daher kann man ohne weiteres von einer flächendeckenden Problematik sprechen.

Falls Sie zu den eiligen Lesern gehören, können Sie Abschnitt 9 auch überfliegen oder ganz überspringen und bei Abschnitt 10 weiterlesen.

9 Und was kam heraus – für den geduldigen Leser

9.1 Analyse ohne alphabetische Effekte

Bevor wir das Gesamtergebnis jedes Tests in Abschnitt 9.1.2 präsentieren, betrachten wir zuerst im folgenden Abschnitt 9.1.1 den interessantesten Teil jedes Rankings, nämlich die Bestenliste.

9.1.1 Die Top 5+ jedes Tests

Für jedes betrachtete Testmedium finden Sie weiter unten in diesem Abschnitt jeweils eine Tabelle. Jedem von uns ausgewählten Einzeltest ist darin wieder eine eigene Zeile gewidmet. Grundidee in diesem Abschnitt ist, jeweils die „Top 5“ bestbewerteten Produkte zu betrachten, auf die sich ja in der Regel die Aufmerksamkeit der Leserinnen und Leser konzentriert, und die besonders intensiv beworben werden.

Allerdings kommt hier eine Feinheit hinein, weswegen wir von „Top 5+“ sprechen. Das Problem ist, dass punkt-/notengleiche Produkte im Ranking ja denselben bereinigten Platz einnehmen. Zum Beispiel Superzoomkameras bei fotoMagazin: Auf dem ersten bis vierten bereinigten Platz findet sich jeweils genau ein Produkt, auf dem fünften bereinigten Platz aber gleich vier Produkte.[3] Da zwischen diesen vier Produkten keine Reihung möglich ist, werden sie alle vier hineingenommen, so dass insgesamt acht Produkte in den „Top 5+“ sind. Der Extremfall tritt beim Test von Farbschutzschampoos durch Stiftung Warentest auf, wo schon fünfzehn Produkte gemeinsam auf dem ersten bereinigten Platz sind und daher diese fünfzehn Produkte die „Top 5+“ bilden.

In den Tabellen unten ist die Anzahl der insgesamt einbezogenen Produkte für jeden Test in der dritten Spalte aufgeführt. Die zweite Spalte ist eng darauf bezogen. Zum Beispiel beim Test von City-Fahrradhelmen durch Stiftung Warentest: Hier ist die Zahl 4 eingetragen, weil auf den ersten vier bereinigten Plätzen schon insgesamt sechs Produkte platziert sind, so dass die in die Betrachtung hineingenommenen Top 5+ Produkte sich auf nur vier bereinigte Plätze verteilen. Und wieder Extremfall Farbschutzshampoos: Da fünfzehn Produkte auf dem ersten bereinigten Platz sind, werden auch nur Produkte auf einem einzigen bereinigten Platz berücksichtigt (nämlich dem ersten), also steht hier in der zweiten Spalte eine 1.

Bei Betrachtung der Top 5+ ist natürlich besonders interessant, wie viele Produkte aus den Top 5+ herausfallen bzw. wie viele in die Top 5+ hineinkommen, wenn wir die ursprüngliche Gewichtung durch unsere Variation ersetzen. Beachten Sie, dass beide Zahlen unterschiedlich sein können. Diese beiden Zahlen sind in der dritten und vierten Spalte eingetragen.

In den Spalten mit Zahlen als Titel sind die einzelnen Sprungweiten notiert. Genauer gesagt, gibt die für einen Test eingetragene Zahl jeweils an, wie viele Produkte so weit gesprungen sind, wie es die Spaltenüberschrift angibt. Nur tatsächlich vorkommende Sprungweiten sind berücksichtigt, so dass sich eine Lücke in den Zahlen ergibt, wenn eine Sprungweite nicht auftritt. Für diese Spalten haben wir alle Produkte gezählt, die entweder im ursprünglichen Test oder mit unserer Variation der Gewichtungen in den Top 5+ landen (oder beides). Mit anderen Worten: Sowohl

[3] Vergleiche die Tabelle zu Superzoomkameras in Abschnitt 1.1 von „Sind Rankings inhärent willkürlich? – Detaillierte Daten“.

die Sprünge innerhalb der Top 5+ als auch in die Top 5+ hinein als auch aus den Top 5+ heraus werden mitgezählt.

Beispiel:

- Die erste Tabelle unten (Stiftung Warentest) hat die Spalten „1“, „3“, „4“ und „5“. Beim Test von Multivitaminsäften stehen in diesen Spalten die Einträge „1“, „0“, „1“ und „1“. Das bedeutet, dass ein Produkt um einen bereinigten Platz gesprungen ist, ein anderes Produkt um vier bereinigte Plätze und ein weiteres Produkt sogar um fünf bereinigte Plätze. Die „2“ ist keine Spaltenüberschrift, da in keinem einzigen Test bei Stiftung Warentest Sprungweite 2 auftrat.

Die Spalte „Summe Sprungweiten“ summiert die Sprungweiten aller berücksichtigten Produkte auf, so dass etwa bei Multivitaminsäften 1+4+5=10 herauskommt.

Wie schon in Abschnitt 7.2 ausgeführt, betrachten wir in den letzten beiden Spalten die Anzahl der Produktpaare, die im originalen Test gleich bewertet werden und in unserem Test mit variierten Gewichtungen nicht („Vorher gleich bewertet“) bzw. die in unserem Test gleich bewertet werden und im originalen Test nicht („Vorher ungleich bewertet“).

Stiftung Warentest

Test	Anzahl verschiedene Noten Top 5+ im originalen Ranking	Anzahl Produkte in Top 5+ im originalen Ranking	Abstiege	Aufstiege	1	3	4	5	Summe Sprungweiten	Vorher gleich bewertet	Vorher ungleich bewertet
Rote Nagellacke	4	7	0	1	0	1	0	0	3	1	3
Smartphones	2	6	1	0	1	1	0	0	4	4	0
Wandfarben	2	5	0	1	1	1	0	0	4	3	1
City-Fahrradhelme	4	6	0	0	3	0	0	0	3	1	2
Farbschutzshampoo	1	15	0	0	0	0	0	0	0	0	0
Drucker-Scanner	4	5	0	0	2	0	0	0	2	0	2
Bodenstaubsauger	4	6	1	0	2	0	0	0	2	2	0
Elektrovertikutierer	3	5	0	0	3	0	0	0	3	2	1
Multivitaminsaft	4	9	1	0	0	0	0	1	5	5	0
Sommerreifen	3	6	0	0	3	0	0	0	3	2	1

Es zeigen sich doch einige Sprünge um drei oder mehr Positionen. In fünf der neun Tests ändert sich die Menge der Top 5+ Produkte.

Besonders interessant sind natürlich Bewegungen beim ersten Platz:

- *Smartphones*: Motorola Razar steigt vom (geteilten) ersten auf den zweiten Platz ab.
- *Sommerreifen*: Maloya Lugano steigt vom zweiten auf den (geteilten) ersten Platz auf.

Nur bei Farbschutzschampoos ändert sich überhaupt nichts in den Top 5+. Allerdings sind fünfzehn Produkte gemeinsam auf dem ersten Platz, so dass die Top 5+ in diesem Fall sowieso keine Aussagekraft haben.

CHIP Online

Test	Anzahl verschiedene Punkte Top 5+ im originalen Ranking	Anzahl Produkte in Top 5+ im originalen Ranking	Abstiege	Aufstiege	1	2	3	5	6	Summe Sprungweiten	Vorher gleich bewertet	Vorher ungleich bewertet
3D-Fernseher 43 Zoll	5	5	1	3	0	1	1	1	1	16	0	3
MP3-Player	5	5	0	0	0	0	0	0	0	0	0	0
Notebooks 14-15,6 Zoll	4	5	1	1	0	2	2	0	0	10	1	1
Tablets	5	5	1	1	1	2	0	0	0	5	0	1

Besonders auffällig bei den Top 5+ ist der Test der 3D-Fernseher mit je einem Sprung um zwei, drei, fünf bzw. sechs Positionen sowie drei Aufstiegen in die Top 5+.

ETM TESTMAGAZIN

Test	Anzahl verschiedene Punkte Top 5+ im originalen Ranking	Anzahl Produkte in Top 5+ im originalen Ranking	Abstiege	Aufstiege	1	Summe	Vorher gleich bewertet	Vorher ungleich bewertet
Kochtöpfe	5	5	0	0	3	3	0	1
Standmi-xer	5	5	0	0	3	3	0	1
Dörrgeräte	5	5	0	1	1	1	0	1

Bei den Standmixern wechseln die ersten beiden Produkte ihre Positionen (KitchenAid artisan 5KSB555 wird von Gastroback Design Mixer Advanced Electronic 41001 überholt). Bei den Kochtöpfen wechseln die Produkte auf Platz 2 und 3 ihre Positionen (Fissler Iuno und Woll Concept pro).

Unsere Ergebnisse für *Haus & Garten Test* und für *fotoMagazin* präsentieren wir im Folgenden kommentarlos; mit den bisherigen Ausführungen sollten sie auch ohne Kommentare unsererseits problemlos zu interpretieren sein.

Haus & Garten Test

Test	Anzahl verschiedene Noten Top 5+ im originalen Ranking	Anzahl Produkte in Top 5+ im originalen Ranking	Abstiege	Aufstiege	1	2	Summe Sprungweiten	Vorher gleich bewertet	Vorher ungleich bewertet
Vakumierer	4	6	1	0	1	1	3	1	2
Kohlegrills	5	5	0	0	1	0	1	0	1
Gasgrills	4	6	1	0	3	0	3	2	1

fotoMagazin

Test	Anzahl verschiedene Punkte Top 5+ im originalen Ranking	Anzahl Produkte in Top 5+ im originalen Ranking	Abstiege	Aufstiege	1	4	Summe Sprungweiten	Vorher gleich bewertet	Vorher ungleich bewertet
Kompaktkameras	4	6	0	0	4	0	4	1	3
Superzoomkameras	5	8	1	1	0	2	8	3	4

Shanghai-Ranking

Zwei Universitäten unter den Top 5 wechseln ihre Plätze: Das MIT fällt vom dritten auf den vierten Platz zurück, Berkeley steigt vom vierten auf den dritten Platz auf. Die genaue Platzierung im Shanghai-Ranking hat für jede führende US-Uni erhebliche Konsequenzen.

Bemerkung: Wie in Abschnitt 7 erläutert, haben wir jedes einzelne Gewicht um maximal einen Prozentpunkt vermindert und um maximal zwei Prozentpunkte erhöht. Wenn wir aber eine doppelte Variationsbreite erlauben, das heißt, Verminderung um maximal zwei Prozentpunkte und Erhöhung um maximal vier Prozentpunkte, dann springt Stanford sogar um drei Positionen

vom zweiten auf den fünften Platz (MIT, Berkeley und Cambridge rutschen entsprechend um eine Position hoch). Für Stanford wäre dies natürlich ein Desaster.

9.1.2 Gesamttests

Bei den Gesamttests macht es Sinn, wie in Abschnitt 8 die durchschnittliche Sprungweite mit zu betrachten. Ansonsten sind die folgenden Tabellen genauso aufgebaut wie die in Abschnitt 9.1.1.

Stiftung Warentest

Test	Anzahl getestete Produkte	Anzahl verschiedene Noten im originalen Ranking	1	2	3	4	5	Summe Sprungweiten	Durchschnittliche Sprungweite	Vorher gleich bewertet	Vorher ungleich bewertet
RoteNagellacke	21	14	3	0	1	0	0	6	0,28	1	5
Smartphones	21	7	3	3	4	0	0	21	1	14	11
Wandfarben	20	9	3	0	1	0	0	6	0,3	4	2
City-Fahrradhelme	16	11	6	0	0	0	0	6	0,37	1	5
Farbschutzshampoo	16	2	0	0	0	0	0	0	0	0	0
Drucker-Scanner	6	5	2	0	0	0	0	2	0,33	0	2
Bodenstaubsauger	10	8	2	0	0	0	0	2	0,2	2	0
Elektrovertikutierer	15	10	3	0	0	0	0	3	0,2	2	1
Multivitaminsaft	20	13	0	0	0	0	1	5	0,25	5	0
Sommerreifen	16	11	3	0	0	0	0	3	0,18	2	1

Beim originalen Testergebnis für Smartphones kommen nur sieben verschiedene Endnoten heraus. Trotz dieser geringen Trennschärfe wechseln immer noch zehn der insgesamt einundzwanzig getesteten Produkte ihre Position.

Bei Multivitaminsäften mit dreizehn verschiedenen Endnoten springt ein Produkt um vier und ein wieteres um fünf Positionen.

Die folgenden Ergebnisse präsentieren wir wieder kommentarlos.

CHIP Online

Test	Anzahl getestete Produkte	Anzahl verschiedene Punkt-zahlen im originalen Ranking	1	2	3	4	5	Summe Sprungweiten	Durchschnittliche Sprungweite	Vorher gleich bewertet	Vorher ungleich bewertet
3D-Fernseher 43 Zoll	28	26	8	4	1	1	2	33	1,17	2	3
MP3-Player	8	8	0	0	0	0	0	0	0	0	0
Notebooks14-15,6 Zoll	17	16	6	3	1	0	0	15	0,88	1	0
Tablets	40	39	11	9	0	0	0	29	0,72	1	4

ETM Testmagazin

Test	Anzahl getestete Produkte	Anzahl verschiedene Punkt-zahlen im originalen Ranking	1	2	3	Summe Sprungweiten	Durchschnittliche Sprungweite	Vorher gleich bewertet	Vorher ungleich bewertet
Kochtöpfe	24	24	7	2	2	17	0,7	0	5
Standmixer	13	13	5	0	0	5	0,38	0	3
Dörrgeräte	7	7	1	0	0	1	0,14	0	1

Haus & Garten Test

Test	Anzahl getestete Produkte	Anzahl verschiedene Noten im originalen Ranking	1	2	Summe Sprungweiten	Durchschnittliche Sprungweite	Vorher gleich bewertet	Vorher ungleich bewertet
Vakuumierer	6	4	1	1	3	0,5	1	2
Kohlegrills	5	5	1	0	1	0,2	0	1
Gasgrills	6	4	3	0	3	0,5	1	2

fotoMagazin

Test	Anzahl der Produkte	Anzahl verschiedene Punkt-zahlen im originalen Ranking	1	2	4	Summe Sprungweiten	Durchschnittliche Sprungweite	Vorher gleich bewertet	Vorher ungleich bewertet
Kompaktkameras	20	12	9	1	0	11	0,55	5	6
Superzoomkameras	16	9	6	1	2	16	1	7	7

Shanghai-Ranking

Zusätzlich zum Austausch von MIT und Berkeley unter den Top 5, den wir schon in Abschnitt 9.2.1 erwähnt hatten, haben noch zwei weitere hochrangige amerikanische Universitäten ihre Plätze im Ranking getauscht: Penn State und Cornell (Platz 13 und 14).

9.2 Potentiell irreführende alphabetische Sortierung

In Abschnitt 7.3 haben wir unsere Vorgehensweise erläutert, um die Verschiebungen in Rankingtabellen zu messen, in denen gleichplatzierte Produkte in alphabetischer Reihenfolge aufgeführt werden. Wie schon am Ende von Abschnitt 7.3 gesagt, betrifft dies unter den von uns betrachteten Medien nur *Stiftung Warentest*, *CHIP Online* und *fotoMagazin*.

Parallel zu Abschnitt 9.1 konzentrieren wir uns in Abschnitt 9.2.1 erst einmal auf die ersten fünf Plätze bei jedem Test, bevor wir in Abschnitt 9.2.2 die Gesamttests betrachten.

9.2.1 Die Top 5 jedes Tests

Bei alphabetischer Sortierung punkt-/notengleicher Produkte ist jeder einzelne Platz in der Reihung durch genau ein Produkt besetzt. Daher brauchen wir hier keine Top 5+ zu betrachten, sondern es gibt immer *die* Top 5 Produkte auf den Plätzen 1...5. Beachten Sie, dass daher die Anzahl der Aufstiege in die Top 5 gleich der Anzahl der Abstiege aus den Top 5 ist.

Im Prinzip sind die Tabellen so aufgebaut wie in Abschnitt 9.1.1. Die einzige Ausnahme ist, dass die Anzahl Aufstiege in die Top 5 und die Anzahl Abstiege aus den Top 5 nicht mehr getrennt ausgewiesen werden müssen, sondern – da identisch – unter dem Spaltentitel „Wechsel" zusammengefasst sind. Analog zu Abschnitt 9.1.1 werden in den Spalten mit Ziffern als Überschriften auch hier alle Sprünge aus den Top 5 heraus und in die Top 5 hinein gezählt.

Stiftung Warentest

Test	Wechsel	1	2	3	5	Summe der Sprungweiten
RoteNagellacke	0	0	0	0	0	0
Smartphones	1	5	0	1	0	8
Wandfarben	1	1	2	0	0	5
City-Fahrradhelme	0	2	0	0	0	2
Farbschutzshampoo	0	0	0	0	0	0
Drucker-Scanner	0	2	1	0	0	4
Bodenstaubsauger	1	2	0	0	0	2
Elektrovertikutierer	0	2	0	0	0	2
Multivitaminsaft	1	2	0	0	1	7
Sommerreifen	0	2	1	0	0	4

In vier von insgesamt zehn Tests hat sich nicht nur die Reihenfolge, sondern auch die Zusammensetzung der Top 5 geändert. Bei Smartphones wechseln Platz 1 und 2 (Motorola Razar wird von Samsung Galaxy Note überholt). Bei Farbschutzshampoos ist die Situation wieder so undifferenziert, dass sich nichts ändern kann.

CHIP Online

Test	Wechsel	1	2	3	4	5	6	Summe
3D-Fernseher43Zoll	1	2	0	0	0	1	1	13
MP3-Player	0	0	0	0	0	0	0	0
Notebooks14-15,6Zoll	1	1	3	0	1	0	0	11
Tablets	1	2	1	0	0	0	0	4

Hier gab es bei drei von vier Tests je einen Wechsel unter den Top 5. Bemerkenswert ist beim Test der 3D-Fernseher, dass das Produkt, das in die Top 5 hineinkommt (Sony KDL-46HX855), dabei um fünf Positionen gesprungen ist, und das Produkt, das aus den Top 5 herausfällt (Panasonic TX-P50VT50E), sogar um sechs Positionen.

fotoMagazin

Test	Wechsel	1	2	Summe
Kompaktkameras	0	2	1	4
Superzoomkameras	0	2	0	2

Leicht auffällig ist der Test der Kompaktkameras mit einem Zweier- und zwei Einersprüngen.

9.2.2 Gesamttests

Schlussendlich stellen wir nun unsere Gesamtergebnisse unter alphabetischen Effekten vor.

Im Prinzip ist das Format der Tabellen dasselbe wie in Abschnitt 9.1.2. Da unter alphabetischer Reihung keine zwei Produkte auf derselben Position landen können, fallen die beiden Spalten „Vorher gleich bewertet“ und „Vorher ungleich bewertet“ weg, die in Abschnitt 9.1.2 noch von Bedeutung waren.

Stiftung Warentest

Test	Anzahl getestete Produkte	1	2	3	5	Summe Sprungweiten	Durchschnittliche Sprungweite
RoteNagellacke	21	2	1	0	0	4	0,19
Smartphones	21	11	3	3	0	26	1,23
Wandfarben	20	2	2	0	0	6	0,3
City-Fahrradhelme	16	5	0	1	0	8	0,5
Farbschutzshampoo	16	0	0	0	0	0	0
Drucker-Scanner	6	2	1	0	0	4	0,66
Bodenstaubsauger	10	2	0	0	0	2	0,2
Elektrovertikutierer	15	2	0	0	0	2	0,13
Multivitaminsaft	20	5	0	0	1	10	0,5
Sommerreifen	16	3	0	1	0	6	0,37

Bei Sommerreifen springt Nokian V vom vierten auf den ersten Platz.

CHIP Online

Test	Anzahl getestete Produkte	1	2	3	4	5	Summe Sprungweiten	Durchschnittliche Sprungweite
3D-Fernseher43Zoll	28	9	3	1	1	2	32	1,14
MP3-Player	8	0	0	0	0	0	0	0
Notebooks14-15,6Zoll	17	7	2	1	0	0	14	0,82
Tablets	40	13	6	1	0	0	28	0,7

Auch hier sind die 3D-Fernseher besonders auffällig mit sechzehn von insgesamt 28 Produkten, deren Positionen sich ändern, davon immerhin bei zweien ein Sprung um fünf Positionen und bei einem um vier. Auch hier springt wieder jedes Produkt im Durchschnitt um mehr als eine Position.

fotoMagazin

Test	Anzahl getestete Produkte	1	2	3	Summe Sprungweiten	Durchschnittliche Sprungweite
Kompaktkameras	20	7	1	1	12	0.6
Superzoomkameras	16	2	4	2	16	1.0

10 CHE-Hochschulranking

Die Einzeltabellen zu den Ergebnissen dieses Abschnitts finden Sie in Abschnitt 1.5 und 1.6 von „Sind Rankings inhärent willkürlich? – Detaillierte Daten“.

Für jede wissenschaftliche Disziplin werden die Fachbereiche aller Hochschulen miteinander verglichen, wobei Fachbereiche an Universitäten und Fachbereiche an Fachhochschulen jeweils unter sich verglichen werden. Wir haben sieben Rankings des Jahres 2012 betrachtet: die universitären Fachbereiche in den Disziplinen Anglistik, Biologie, Chemie, Geographie, Kommunikationswissenschaft, Wirtschaftsingenieurwesen und Zahnmedizin. Unsere Basis ist die Aufbereitung für DIE ZEIT, für Details siehe die Webseite ranking.zeit.de/che2012/de/.

Bei diesen Rankings wurden fünf Kriterien erhoben. In einem raffinierten Verfahren, dessen Details für unsere Untersuchung keine Rolle spielen und daher hier ausgelassen werden,[4] wird für

[4]Beschreibung im Detail: www.che-ranking.de/methodenwiki/index.php/Vorgehensweise.

jeden Fachbereich in jedem der fünf Kriterien ein Punkt vergeben, der grün, gelb oder rot ist. Ein grüner Punkt für ein Kriterium zeigt an, dass der Fachbereich in diesem Kriterium nach CHE-Definition zur Spitzengruppe gehört, bei einem roten Punkt gehört der Fachbereich in diesem Kriterium zur Schlussgruppe, und Fachbereiche mit gelben Punkten sind im Mittelfeld.

Die Reihung der Fachbereiche kommt nun so zustande: Ein Fachbereich schneidet umso besser ab, je mehr grüne Punkte er hat. Haben zwei Fachbereiche dieselbe Zahl an grünen Punkten, dann ist der Fachbereich mit der kleineren Zahl an roten Punkten der höher platzierte von beiden. Haben zwei Fachbereiche dieselbe Anzahl von roten und dieselbe Anzahl von grünen Kriterien (und somit auch dieselbe Anzahl von gelben Kriterien), dann landen sie auf demselben Platz.

In dieser Vorgehensweise liegt eine Asymmetrie, an der wir ansetzen:

- *Umkehrung*: Statt zuerst die grünen Punkte positiv zu zählen und nur bei Gleichstand die roten Punkte zu betrachten, könnte man genauso gut die Asymmetrie umdrehen, also zuerst die roten Punkte negativ zählen und bei Gleichstand die grünen Punkte betrachten.
- *Differenzbildung*: Um die Asymmetrie zwischen grün und rot zu vermeiden, könnte man sich auch „in der Mitte treffen" und einfach die Anzahl der roten Punkte von der Anzahl der grünen Punkte abziehen und diese Differenz entscheiden lassen.

Beides haben wir gemacht: die Asymmetrie umdrehen bzw. „sich in der Mitte treffen". In der folgenden Tabelle sind für beide Varianten jeweils sämtliche einzelnen Sprungweiten aufgeführt analog zur Darstellung in Abschnitt 8.

Fachbereiche, die nicht in jedem Kriterium eine Bewertung haben, werden von uns nicht betrachtet.

Ohne alphabetische Sortierung (gemäß Abschnitt 7.2):

Disziplin	Anzahl Fachbereiche	Sprünge bei Umkehrung	Sprünge bei Differenzbildung
Anglistik	27	1x11, 1x10, 1x8, 1x7, 1x4, 6x3, 3x2, 4x1	2x6, 1x3, 8x2, 7x1
Biologie	40	1x12, 6x8, 2x6, 5x4, 2x3, 3x2, 4x1	3x4, 7x3, 3x2, 4x1
Chemie	35	1x18, 1x15, 2x11, 1x9, 1x8, 2x7, 6x5, 4x4, 2x2, 2x1	2x8, 1x7, 1x4, 9x3, 5x2, 7x1
Geografie	20	2x7, 1x3, 6x2, 1x1	1x2, 1x1
Komm.wiss.	11	1x6, 1x5, 2x3, 1x2, 1x1	1x4, 1x2, 5x1
Wirtsch.ing.	19	2x6, 2x5, 7x2	2x3, 3x2
Zahnmedizin	12	1x6, 3x2, 4x1	2x2, 2x1

Mit alphabetischer Sortierung (gemäß Abschnitt 7.3):

Disziplin	Anzahl Fachbereiche	Sprünge bei Umkehrung	Sprünge bei Differenzbildung
Anglistik	27	1x11, 1x10, 1x8, 1x7, 1x4, 6x3, 3x2, 4x1	1x8, 1x5, 1x3, 4x2, 10x1
Biologie	40	1x12, 6x8, 2x6, 5x4, 2x3, 3x2,	1x6, 2x4, 2x3, 3x2,

		4x1	6x1
Chemie	35	1x18, 1x15, 2x11, 1x9, 1x8, 2x7, 6x5, 4x4, 2x2, 2x1	2x8, 2x7, 3x4, 2x3, 7x2, 8x1
Geografie	20	2x7, 1x3, 6x2, 1x1	2x1
Komm.wiss.	11	1x6, 1x5, 2x3, 1x2, 1x1	1x2, 6x1
Wirtsch.ing.	19	2x6, 2x5, 7x2	2x2, 4x1
Zahnmedizin	12	1x6, 3x2, 4x1	2x2, 2x1

In beiden Fällen ist das typische Bild, dass eine kleinere Zahl von Fachbereichen größere Sprünge nach unten machen, und die „übersprungenen“ Fachbereiche rücken entsprechend jeweils um eine Position auf. Wie zu erwarten, sind diese Effekte bei der Umkehrung wesentlich größer als bei der Differenzbildung. Große Sprünge sind nicht allzu häufig, dann aber durchaus weit genug, um Fachbereiche aus der Spitzengruppe in das untere Mittelfeld bzw. aus dem oberen Mittelfeld in die Schlussgruppe abstürzen zu lassen.

Tatsächlich sind große Sprünge nur von guten zu schlechteren Plätzen zu erwarten – nicht etwa umgekehrt –, da das CHE-Ranking Zweifelsfälle deutlich positiv bewertet, während solche Zweifelsfälle bei genau umgedrehter Logik entsprechend deutlich negativ bewertet werden.

11 Ökotest

Entgegen allgemeiner Wahrnehmung erstellt Ökotest nicht unbedingt Rankings, sondern verfolgt oft andere Zielsetzungen. Zum Verständnis eines Testergebnisses von Ökotest ist unbedingt wichtig, die Rubrik „Bewertung“ im Kasten „So haben wir getestet“ zu lesen. Hier sollte die Zielsetzung des Tests zu finden sein. Zum Beispiel liest man zum Test „Reiseapotheke“ im Heft 06/2012 auf Seite 56: „Ziel unseres Tests war es, Ihnen eine Reiseapotheke für den Urlaub empfehlen zu können, die ausschließlich ‚sehr gute´ und ‚gute´ Präparate enthält ...“ Allerdings muss wohl davon ausgegangen werden, dass Tests von Ökotest entgegen ihrer ausdrücklichen Intention doch häufig als Rankings missverstanden werden. So werben nicht wenige Hersteller mit Formulierungen wie „Testsieger Ökotest“.

Die Tests, die wir uns angeschaut haben, sind eher grobgranular im Vergleich zu den Tests in anderen Medien und bieten daher nur eine grobe Orientierung. Nehmen wir etwa den Test Laktose freier Produkte aus Heft 06/2012, Seite 35 ff., oder etwa den Test von Deos mit Langzeitwirkung aus demselben Heft, Seite 84 ff.: Nur zwei Teilnoten fließen ein, die aber nicht auf gleicher Stufe miteinander verrechnet werden, sondern die erste Teilnote ist die eigentliche Note, aber aufgrund der zweiten Teilnote kann es Abzüge bei der eigentlichen Note geben. Damit ist nur eine sehr grobe Unterscheidung zwischen Produkten möglich, und es gibt daher auch keinen Angriffspunkt für unseren Ansatz. Das ist natürlich uneingeschränkt positiv, denn wie wir in diesem Aufsatz ja zeigen, ist mehr als eine solche grobe Orientierung wohl nicht seriös möglich.

Allerdings gibt es auch bei Tests dieser Art durchaus noch eine Angriffsfläche. Wenn zum Beispiel das Vorhandensein eines Schadstoffes einen Abzug um gleich zwei Noten nach sich zieht, kann es schon eine Rolle spielen, wie hoch der Grenzwert für den Schadstoff gesetzt wird, ab dem der Notenabzug passiert. Nach eigener Aussage verwendet Ökotest selbstgewählte schärfere Grenzwerte als die gesetzlichen. Im Geiste dieses Aufsatzes wäre es sinnvoll nachzuprüfen, welche Effekte ein kleines „Ruckeln“ an diesen Grenzen hätte. Dafür müssten allerdings die Rohdaten für die Notenbildung zur Verfügung gestellt werden. Eine entsprechende Anfrage an Ökotest per E-Mail blieb leider ohne Antwort.

Auch Ökotest stellt bisweilen Rankings auf, beispielsweise der viel beachtete Strompreisrechner aus Heft 03/2008, Rubrik Geld und Recht. Leider haben wir nicht einmal die Berechnungsmethode gefunden. Es wird nur auf Seite 166 gesagt, dass es 30 Punkte bringt, wenn der Strompreisrechner den günstigsten Tarif auflistet, 20 Punkte beim zweitgünstigsten und ein Punkt beim dreißigst günstigsten – aber die genaue Berechnungsmethode, mit der ein Anbieter auf maximal 465 Punkte kommen kann, wird nicht publiziert. Ebenso fehlen die Rohdaten, welcher Strompreisrechner nun welche Tarife gefunden hat und welche nicht, so dass unsere Methode mangels Transparenz nicht anwendbar ist.

12 Resümee

Wie diese Studie zeigt, ist unsere Fragestellung hoch relevant. In der Tat stellt sich die dringliche Frage, ob die Veröffentlichung von – und Werbung mit – Endnoten bzw. Platzierungen in Rankings überhaupt seriös sein *kann*. Wie wir auch gesehen haben, löst der Übergang auf Schulnoten das Problem nicht, sondern produziert neue Probleme.

Eine Möglichkeit wäre, die Einzelnoten *nicht* zu einer Gesamtnote zusammenzurechnen, sondern als Einzelnoten stehen zu lassen. Dies wäre ein Kulturwechsel beim Umgang der interessierten Öffentlichkeit mit Tests und Studien – weg vom sportlichen Wettkampf um „die Medaille“, hin zu einer Kultur, in der sich jeder Konsument die einzelnen Kriterien genau anschaut und selbstverantwortlich entscheidet, welche Kriterien ihm zu welchem Grad wichtig sind.

Natürlich ändern alle diese Überlegungen aber nichts daran, dass es jenseits unseres rein mathematischen Fokus’ auch fachlich begründete Zweifel an den einzelnen Kriterien der verschiedenen Rankings geben kann.