Dr. Daniel Vorgrimler, Dipl.-Ing. agr. Dirk Wübben[1])

Die Delphi-Methode und ihre Eignung als Prognoseinstrument

Das Orakel von Delphi war in der Antike eine der bedeutendsten Kultstätten und eine wichtige politische Entscheidungshilfe. Im sakralen Bereich des Tempels von Delphi auf einem Dreifuß über einer Erdspalte sitzend erteilte Pythia, die hohe Priesterin Apollos, ihre dunklen und geheimnisvollen Weissagungen, die ihr der Sage nach im Zustand der Trance von Apollo selbst übermittelt wurden. Da die Prophezeiungen der Pythia häufig recht vage waren, kamen die Ratsuchenden meist mehrfach, um das Orakel zu befragen.

Ähnlich verläuft heute eine Experten-Befragung nach der Delphi-Methode. Die Delphi-Methode, ein Instrument der Prognostik, wurde in den 1960er-Jahren zur Voraussage technischer, wirtschaftlicher und sozialer Entwicklungen konzipiert. Es handelt sich um eine mehrstufige Befragung, welche unter Experten verschiedener Fachbereiche schriftlich durchgeführt wird. Dabei wird davon ausgegangen, dass Experten in ihrem Fachgebiet über besonders viel Wissen verfügen und deshalb sehr gute Schätzungen über mögliche zukünftige Entwicklungen abgeben können. Durch Rückkopplung der Zwischenergebnisse an die Befragten wird den Teilnehmern eine Möglichkeit zur Überprüfung bzw. eines Vergleichs ihrer Aussagen mit den Meinungen der anderen Experten gegeben. Durch die wiederholte Befragung soll die Spannweite der Expertenmeinungen verringert werden. Es geht bei diesem qualitativen Prognoseverfahren also weniger um die Abbildung quantitativer Meinungsverhältnisse, als darum, Inhalte zu verdichten und damit besonders aussagekräftige Prognosen aufstellen zu können.

In der amtlichen Statistik werden intuitive Prognoseverfahren wie die Delphi-Methode nicht angewandt. Ungeachtet dessen werden neuere Entwicklungen auf dem Gebiet der statistischen Methoden und Befragungstechniken aufmerksam verfolgt, wie auch der nachfolgende Beitrag zeigt.

Er beleuchtet in einem ersten Teil die Vorgehensweise einer Delphi-Befragung, macht die Vor- und Nachteile sichtbar und zeigt, welche „Prozesse“ sich die Befürworter dieser Befragungstechnik erwarten. Im zweiten Teil wird beispielhaft für eine durchgeführte Delphi-Befragung analysiert, inwieweit eine Delphi-Befragung die in sie gestellten Erwartungen erfüllen kann.

Vorbemerkung

Prognosen können mittels verschiedener Vorgehensweisen erstellt werden. Neben quantitativen Verfahren wird auch auf qualitative bzw. intuitive Verfahren zurückgegriffen. Quantitative Methoden werden vorwiegend bei so genannten „sicheren Aussagen“ angewandt. Eine Aussage gilt dann als sicher, wenn (Natur-)Gesetze zum Prognosethema existieren (z. B. für die Vorhersage der Planetenbewegungen). Wenn breite empirische Kenntnisse die Entwicklung eines mathematischen Modells ermöglichen, kann von einem relativ geringen Schwierigkeitsgrad bei der Prognose ausgegangen werden. Je weniger hingegen ein Prognoseobjekt theoretisch oder empirisch fundiert ist, desto höher ist der Schwierigkeitsgrad der Prognoseerstellung

1) Daniel Vorgrimler, der mittlerweile im Statistischen Bundesamt arbeitet, war wissenschaftlicher Mitarbeiter am Institut für Agrarpolitik und landwirtschaftliche Marktlehre der Universität Hohenheim; Dirk Wübben war Diplomand am gleichen Institut.

und desto zweckmäßiger ist die Anwendung qualitativer Prognosemethoden.[2])

Insbesondere dort, wo Entwicklungen weder direkt noch indirekt aus der Vergangenheitsentwicklung ableitbar sind, spielen intuitive Prognoseverfahren eine Rolle. Sie sind dadurch gekennzeichnet, dass die Theorien, auf welchen die Prognosen aufbauen, mit subjektiven und nicht unmittelbar nachprüfbaren Zukunftseinschätzungen infiltriert sind. Insbesondere „langfristige Prognosen müssen sich auch auf Prognosen von Strukturbrüchen erstrecken, sie müssen versuchen, nicht nur den Wertbereich bestimmter Ereignisse vorauszusagen, sondern auch grundsätzlich neue Ereignisse als möglich und wünschenswert erkennen“[3]).

Die Expertenbefragung nach der Delphi-Methode stellt das gebräuchlichste intuitive Prognoseverfahren dar. Im folgenden Kapitel wird diese Art der Expertenbefragung theoretisch vorgestellt. Ihre in der Theorie erwarteten Prozesse werden im zweiten Kapitel anhand eines empirischen Beispiels überprüft. Das abschließende Fazit bewertet die Delphi-Methode.

1 Die Delphi-Methode

1.1 Hintergrund und Definition der Delphi-Methode

Bei der Delphi-Methode handelt es sich um ein mehrstufiges Befragungsverfahren mit Rückkopplung. Mehrere Experten beantworten anonym einen Fragebogen zum Prognosethema. Die Ergebnisse werden ausgewertet und den Teilnehmern mitgeteilt (kontrollierte Rückkopplung). Diese sollen die Ergebnisse überdenken, dazu Stellung nehmen und sie eventuell modifizieren. Eine Delphi-Studie besteht aus mindestens zwei Befragungsrunden.

Wenn auch der Literatur einstimmig zu entnehmen ist, dass es keine allgemein anerkannte Definition der Delphi-Methode gibt, kann doch davon ausgegangen werden, dass über die Grundidee der Delphi-Methode Konsens besteht. Bei den Definitionen der Delphi-Methode stehen oftmals unterschiedliche Aspekte im Vordergrund. Für eine Reihe von Autoren sind die gruppendynamischen Eigenschaften zentrales Wesensmerkmal der Methode. Die Delphi-Technik wird von diesen Autoren als allgemeines Instrument zur verbesserten Erfassung von Gruppenmeinungen und zur Steuerung der Gruppenkommunikation angesehen.[4]) Andere Autoren betonen stärker inhaltliche Aspekte, insbesondere den Problemlösungscharakter und den Umgang mit Unsicherheit. Die Delphi-Methode als qualitativer Ansatz zur Erstellung von Prognosen steht hier im Zentrum. Ono/Wedemeyer zum Beispiel bezeichnen die Delphi-Methode als das wichtigste Verfahren für Zukunftsvorhersagen (”cornerstone of future research“)[5]). Dass die Anwendbarkeit der Methode über Prognosezwecke hinausgeht, macht die Definition von Bortz/Döring deutlich: „Es handelt sich (...) um eine hochstrukturierte Gruppenkommunikation, deren Ziel es ist, aus Einzelbeiträgen der an der Kommunikation beteiligten Personen Lösungen für komplexe Probleme zu erarbeiten.“[6])

1.2 Anwendungsbereiche

Die Delphi-Methode findet auf vielen Gebieten Anwendung. Insbesondere für Aussagen über solche Bereiche, die kostenaufwändige Investitionen erfordern (z. B. die Telekommunikation) oder deren Erfolgskriterien schwer einzuschätzen sind (z. B. das Bildungswesen), wird die Delphi-Technik häufig genutzt.

Die Delphi-Methode kann für unterschiedliche Kategorien von Informationen eingesetzt werden, die dabei auch gleichzeitig ermittelt werden können. Das gilt insbesondere für folgende Informationen[7]):

- Quantitätsangaben: In welchem Maße treten Ereignisse ein?
- Qualitätsangaben: Was ist möglich?
- Zeitangaben: Wann treten Ereignisse ein oder werden neue Möglichkeiten realisiert?
- Wahrscheinlichkeitsangaben: Welche Wahrscheinlichkeiten sind den Angaben zu Quantität, Qualität und Zeit zuzuordnen?
- Bewertungen: Sind Entwicklungen oder neue Möglichkeiten unter Berücksichtigung der Auswirkungen wünschenswert?

Der Delphi-Methode kann also eine gewisse Vielseitigkeit in der Zukunftsforschung zugesprochen werden. Neben der Zukunftsforschung (Prognose) können aber auch andere Anwendungszwecke verfolgt werden[8]):

- Entscheidung,
- Problemlösung,
- Planung,
- Verbesserung des Kommunikationspotenzials und
- Politikbeeinflussung.

2) Siehe Gisholt, O.: „Marketing-Prognosen“, Schriftenreihe des Forschungsinstitutes für Absatz und Handel an der Hochschule St. Gallen, Bd. 15, 1976, S. 45 f.
3) Albach, H.: „Informationsgewinnung durch strukturierte Gruppenbefragung. Die Delphi-Methode“ in Zeitschrift für Betriebswirtschaft, Jg. 40, Ergänzungsheft, 1970, S. 14.
4) Siehe z. B. Linstone, H./Turoff, M.: “Introduction” in Linstone, H./Turoff, M. (Hrsg.): “The Delphi Method: Techniques and applications”, Reading, Mass., 1975, S. 3.
5) Siehe Ono, R./Wedemeyer, D.: “Assessing the Validity of the Delphi Technique” in Future 26 (3), 1994, S. 289.
6) Bortz, J./Döring, N.: „Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler“, Berlin 2002, S. 261.
7) Siehe Geschka, H.: „Delphi“ in Bruckmann, G. (Hrsg.): „Langfristige Prognosen“, Würzburg und Wien 1977, S. 38.
8) Siehe Seeger, T.: „Die Delphi-Methode. Expertenbefragung zwischen Prognose und Gruppenmeinungsbildungsprozessen“ in Hochschulsammlung Philosophie: Sozialwissenschaften, Bd. 8, 1979, S. 26.

In diesem Zusammenhang ist zu erwähnen, dass der Einsatz der Methode an bestimmte Voraussetzungen geknüpft werden muss. Die Validität und die Zuverlässigkeit der Delphi-Methode stehen in engem Kontext zu den zu beurteilenden Sachverhalten, sodass nicht von einer beliebig einsetzbaren Universalmethode gesprochen werden kann.[9])

1.3 Merkmale der Delphi-Methode

Folgende Merkmale sind für die Delphi-Methode charakteristisch[10]):

- die Verwendung eines formalisierten Fragebogens,
- die Befragung von Experten,
- die Anonymität der Teilnehmer untereinander,
- die Ermittlung einer statistischen Gruppenantwort,
- die kontrollierte Rückkopplung und
- die (mehrfache) Wiederholung der Befragung.

Dagegen können andere Aspekte je nach Einzelfall variiert werden. Delphi-Studien unterscheiden sich hinsichtlich:

- der Auswahl der Experten,
- des Umfangs der Expertengruppe,
- der (erforderlichen) Anzahl von Befragungsrunden,
- der Gestaltung der Rückkopplung,
- der Erfragung der Selbsteinschätzung der Experten über deren Kompetenz ("self-rating") und
- der Fragetypen.

1.3.1 Auswahl der Experten

Die Auswahl geeigneter Fachleute ist von großer Bedeutung. Experten sollten über Fachwissen, Überblickwissen in Nachbardisziplinen und Kommunikationsbereitschaft verfügen. Das höhere Informationsniveau der Experten resultiert aus ihrer mittelbaren oder unmittelbaren Beschäftigung bzw. Betroffenheit im Bereich des zu untersuchenden Problems.[11]) Fachleute verfügen dabei nicht nur über (technische) Kenntnisse, sondern auch über komplexe Relevanzsysteme.[12]) Die Expertengruppe sollte möglichst breit gestreut sein und eine interdisziplinäre Zusammensetzung aufweisen.[13])

Die Experten werden oftmals auch gebeten, die eigene Expertise bezüglich der jeweiligen Frage einzuschätzen ("self-rating"), da in vielen Fällen die Studien unterschiedliche Fachgebiete oder Themenfelder beinhalten und somit der individuelle Wissensstand zwischen den Fragen variieren kann. Die Berücksichtigung der Selbsteinschätzungen kann auf unterschiedliche Weise erfolgen. So können zum Beispiel die Antwortprofile nach der Selbsteinschätzung gewichtet werden (vollständiges Feedback)[14]) oder ausschließlich solche Antworten einbezogen werden, die in der Selbsteinschätzung die höchsten Kompetenzwerte repräsentieren (partielles Feedback).

Ob eine Zufallsauswahl der Experten erfolgen soll, ist umstritten. Dabei sprechen einige Gründe gegen die Rekrutierung des Expertengremiums mit Hilfe einer Zufallsstichprobe.[15]) So sind Kenntnisse über die Struktur der Grundgesamtheit die Voraussetzung für Zufallsauswahlen. Es ist jedoch praktisch unmöglich, die Grundgesamtheit der potenziell zum jeweiligen Thema befragbaren und kompetenten Experten zu bestimmen und somit ein Auswahlverfahren für die Rekrutierung zu erstellen. Zudem wären bei einer Zufallsauswahl Antwortausfälle schwerwiegender als bei einer bewussten Auswahl, weil die Expertengruppe möglicherweise nicht mehr in der gewünschten Struktur zusammengesetzt wäre.

Fraglich erscheint, ob die teilnehmenden Experten originäre Ideen tatsächlich in einer Befragung preisgeben. Industriefachleute könnten zum Beispiel eventuelle Wettbewerbsvorteile und Wissenschaftler ihre Erstveröffentlichung gefährdet sehen. Außerdem neigen Experten oft dazu, vorsichtige Urteile abzugeben. Die Konsensbildung könnte also zu einer Verstärkung der konservativen Einschätzung führen.

Für Goodman steht ohnehin nicht allein der Expertenstatus als Auswahlkriterium im Vordergrund, sondern auch die tatsächliche Bereitschaft, an der Befragung teilzunehmen: "It would therefore seem to be more appropriate to recruit individuals who have knowledge of a particular topic and who are consequently willing to engage in discussion upon it without the potentially misleading title of 'expert'."[16]) Es wird sogar behauptet: "Don't hire the best expert you can – or even close to the best. Hire the cheapest."[17])

Die optimale Anzahl der Experten hängt im Wesentlichen von der Komplexität der Fragestellungen und der Fachkenntnis der Teilnehmer ab. Bei gering dimensionierten Problemstellungen und hoher Kompetenz der Teilnehmer reicht eine kleinere Expertengruppe meist aus, wohinge-

9) Siehe Häder, M./Häder, S.: „Neuere Entwicklungen bei der Delphi-Methode. Literaturbericht II", ZUMA-Arbeitspapiere 98/05, 1998, S. 7.

10) Siehe Häder, M./Häder, S.: „Die Delphi-Methode als Gegenstand methodischer Forschung" in Häder, M./Häder, S. (Hrsg.): „Die Delphi-Technik in den Sozialwissenschaften", Wiesbaden 2000, S. 15.

11) Siehe Köhler, G.: „Methodik und Problematik einer mehrstufigen Expertenbefragung" in Hoffmeyer-Zlotnik, J. (Hrsg.): „Analyse verbaler Daten", Opladen 1992, S. 319 f.

12) Siehe hierzu Hitzler, R.: „Wissen und Wesen des Experten" in Hitzler, R. et al. (Hrsg): „Expertenwissen", Opladen 1994, S. 25 ff.: Insbesondere durch das Vorhandensein komplexer Relevanzsysteme unterscheidet sich der Experte vom Spezialisten. Der Experte weiß nicht nur, „was er zur praktischen Bewältigung seiner Aufgaben wissen muss", sondern auch, „was die jeweiligen Spezialisten auf dem von ihm vertretenen Wissensgebiet wissen und wie das, was sie wissen, miteinander zusammenhängt."

13) Siehe Welty, G.: "Problems of selecting experts for delphi exercises" in Academy of Management Journal 15 (No. 1), 1972 , S. 121.

14) Siehe Amara, R./Lipinski, A.: "Some views on the Use of Expert Judgment" in Technological and Social Change 3, 1972, S. 288.

15) Siehe Fußnote 9, S. 23.

16) Goodman, C.: "The Delphi technique a critique" in International Journal of Advanced Nursing, No. 12, 1987, S. 732.

17) Parenté, F./Anderson-Parenté, J.: "Delphi Inquiry Systems" in Wright, G./Ayton, P. (Hrsg.): "Judgmental Forecasting", Wiley, Chichester 1987, S. 137.

gen mit wachsender Komplexität (und mit geringerer Expertise der Teilnehmer) der Umfang der Expertengruppe größer sein muss.[18]) Ferner ergibt sich in der Regel ohnehin eine Beschränkung, da die Zahl der Experten begrenzt ist.

1.3.2 Anzahl der Befragungsrunden

Für die sinnvolle Anzahl der Befragungsrunden gibt es keinen Standard. Die optimale Anzahl der Runden wird theoretisch an einem Abbruchkriterium festgemacht. Über die Definition eines Abbruchkriteriums existieren unterschiedliche Auffassungen. Das bei der Delphi-Methode angestrebte Ziel kann entweder wie bei Scheibe et al. die Erreichung einer hinreichenden Stabilität der Expertenmeinungen sein oder wie bei Richey et al. der Konsens zwischen den Experten über ein zu lösendes Problem.[19]) Insbesondere das Konsenskriterium sollte nicht als allgemeines Ziel angesehen werden, da der Erfolg einer Delphi-Studie nicht generell in einer geringen finalen Streuung der Expertenmeinungen gesehen werden kann. So kann auch die Ermittlung bestehender Divergenzen der Expertenurteile als Erfolg gewertet werden: "For example, a bimodal distribution may occur which will not be registered as a consensus, but indicates an important and apparently insoluble cleft of opinion."[20])

In der praktischen Anwendung sind diese Überlegungen weniger von Bedeutung. Die Anzahl der Befragungsrunden hängt zumeist nicht von einem definierten Abbruchkriterium ab, vielmehr werden durch zeitliche und finanzielle Budgets Grenzen gesetzt. Ferner sinkt mit zunehmender Anzahl der Befragungsrunden die Motivation der Teilnehmer, was zu hohen Panel-Mortalitäten führt. Allgemein kann als Optimum eine minimale Anzahl von Runden bei einem annehmbaren Maß an erzielter Genauigkeit gelten. Zeigt sich bereits nach der ersten Runde, dass zu bestimmten Fragen ein Konsens besteht (frühe Mehrheitsbildung), so bietet es sich an, diese aus dem Frageprogramm zu entfernen. Das Frageprogramm kann schrittweise auf die von den Teilnehmern divergent eingeschätzten Aspekte reduziert werden.

Ferner empfiehlt es sich insbesondere bei qualitativen Fragestellungen, die erste Runde als offene Explorationsrunde anzulegen, was im Extremfall die Versendung von unbeschriebenen Blättern bedeuten würde. So können die Teilnehmer zum Beispiel die spezifischen Ereignisse selbst benennen, ohne dass eine Einflussnahme durch die Delphi-Moderatoren erfolgt.[21]) In den folgenden Runden können dann feste Kategorien vorgegeben werden.

1.3.3 Panel-Mortalität

Die bereits erwähnte Panel-Mortalität ist ein Problem, das bei mehrstufigen Befragungen von großer Bedeutung ist. Bei einer Delphi-Befragung gilt es, die teilnehmenden Experten über einen längeren Zeitraum zur Mitarbeit zu motivieren. Mit einem optimalen Ergebnis ist nur zu rechnen, wenn das Wissenspotenzial in jeder Runde möglichst vollständig zur Verfügung steht und dementsprechend in die Rückmeldungen einfließt.[22])

Antwortausfälle sind vor allem dann kritisch, wenn die Expertengruppe gezielt aus Angehörigen verschiedener Berufsgruppen rekrutiert wurde. In diesem Fall empfiehlt es sich zu beobachten, wie sich die Struktur des Expertenpanels der nicht mehr mitwirkenden Befragten verändert.[23])

1.3.4 Gestaltung der Rückmeldung an die Teilnehmer

Die Vorgehensweise bei der Gestaltung der Rückmeldung an die Teilnehmer ist unterschiedlich. Üblicherweise werden Mittelwerte (in der Regel Median oder arithmetisches Mittel) und in vielen Fällen auch Streuungsmaße zurückgemeldet, gelegentlich aber auch graphische Darstellungen, Tabellen oder verbale Kommentare. Des Weiteren werden Experten mit besonders großen Abweichungen meist darum gebeten, die Gründe für ihre extremen Ansichten anzugeben, um den Informationsgehalt zu maximieren.

Problematisch bei der Rückinformation der Teilnehmer ist, dass eine völlige Objektivität in der Projektdurchführung kaum möglich ist. Das Projektteam muss für das Feed-back Zusammenfassungen, Neuformulierungen und Auswahlentscheidungen vornehmen, wodurch Beeinflussungen im Meinungsbildungsprozess der Teilnehmer nicht auszuschließen sind. In keinem Fall darf die Delphi-Moderation eigene (wirtschaftliche) Interessen an bestimmten Ergebnissen einer Delphi-Studie haben, damit nicht die Gefahr einer bewussten oder unbewussten Manipulation der Ergebnisse durch die Rückmeldung besteht.

1.3.5 Anonymität

Wesentliches Element der Delphi-Methode ist die Anonymität der Experten untereinander. Der Meinungsaustausch erfolgt über die Delphi-Moderatoren. Die Anonymität trägt erstens dazu bei, eine Meinungsführerschaft (z. B. durch Dominanz einzelner Personen) in der Expertengruppe zu verhindern. Zweitens liegt es nahe, dass durch eine anonyme Erhebungssituation die Bereitschaft zur Beteiligung an einer Befragung, in der es darum geht, unter Unsicherheit ein Urteil abzugeben, erhöht wird. Drittens werden die Teilnehmer in der anonymen Erhebungssituation vor einem möglicherweise bei einer Meinungsänderung zu befürchtenden Prestigeverlust geschützt.[24]) Schließlich entsteht viertens keine irrelevante Kommunikation, was bei offenen Gruppendiskussionen der Fall sein kann. Insgesamt wird durch die Anonymität bei der Delphi-Methode das Auftreten von sozio-psychologischen Effekten, welche die Mei-

18) Siehe Fußnote 9, S. 24 f.
19) Siehe Scheibe, M. et al.: "Experiments in Delphi Methodology" in Linstone, H./Turoff, M. (Hrsg.), a.a.O., S. 277 f.
20) Siehe Fußnote 19, S. 277.
21) Siehe Fußnote 8, S. 89 f.
22) Siehe Fußnote 10, S. 19.
23) Siehe Williams, P./Webb, C.: "The Delphi technique: a methodological discussion" in Journal of Advanced Nursing 19, 1994, S. 184.
24) Siehe Fußnote 10, S. 17.

nungsbildung in Gruppen oft verzerren, auf ein Minimum reduziert.

Die Anonymität kann sich jedoch auch als Nachteil erweisen, da die Experten für ihre Einschätzungen nicht verantwortlich gemacht werden können. Dabei ist auch nicht auszuschließen, dass sie durch die Anonymität zu einer übereilten, nicht ausreichend durchdachten Urteilsfindung neigen.[25]) Des Weiteren wird kritisiert, dass ein Lernprozess, der durch die Auseinandersetzung mit fremden Argumenten manchmal hervorgerufen wird, durch die Anonymität nicht ausreichend stattfinden kann.[26])

1.4 Evaluation der Delphi-Methode

1.4.1 Theoretische Fundierung der Delphi-Methode

Eine Evaluation der Delphi-Methode ist notwendig, um deren Legitimation als Prognoseinstrument zu prüfen. Grundsätzlich gilt für Verfahren der Informationsgewinnung, dass die Ergebnisse die Gütekriterien Reliabilität (Zuverlässigkeit) und Validität (Gültigkeit) erfüllen müssen.[27])

Eine Befragungsmethode gilt dann als zuverlässig, wenn eine Wiederholung unter gleichen Bedingungen das gleiche Ergebnis erzielt.[28]) Sackman bezweifelt die Reliabilität der Delphi-Methode, weil eine Wiederholung der Befragung wegen der Dynamik der durch die Methode gewonnenen subjektiven Einstellungen zu den zu prognostizierenden Ereignissen keinen Sinn mache.[29]) Tatsächlich lässt sich die Reliabilität der Delphi-Methode nicht exakt bestimmen. Realisierbar wäre ein so genanntes Retest-Verfahren, indem dieselbe Teilnehmergruppe zum selben Vorhersage- bzw. Beurteilungsgegenstand wiederholt befragt werden würde. Problematisch ist allerdings, dass dieses Verfahren sowohl zu einer Überschätzung als auch zu einer Unterschätzung der Reliabilität führen könnte. Die Reliabilität wird überschätzt, wenn die Urteile der Vorbefragung lediglich wiederholt werden. Sie wird unterschätzt, wenn die Urteile aufgrund zwischenzeitlicher Lernprozesse anders ausfallen. Sackmans Kritik greift deshalb nicht hinsichtlich der eigentlichen Reliabilität der Delphi-Methode, sondern hinsichtlich deren exakter Bestimmung. Wechsler empfiehlt zur Überprüfung der Reliabilität Untersuchungen, in denen von verschiedenen Expertengruppen zum selben Zeitpunkt erstellte Vorhersagen verglichen werden.[30]) So wird die Reliabilität zwar weniger im Sinne einer experimentellen Reproduzierbarkeit, dafür aber vielmehr im Sinne einer informationellen Reproduzierbarkeit[31]) ermittelt. Auf diese Weise könne ein empirischer Nachweis der Reliabilität der Delphi-Methode erbracht werden.

Das Kriterium Validität bezieht sich auf die Frage, ob tatsächlich das erhoben wird, was ermittelt werden soll.[32]) Sackman bezweifelt die Validität der mittels der Delphi-Methode abgeleiteten Ergebnisse, weil sie lediglich subjektive Einstellungen bezüglich der zu prognostizierenden zukünftigen Ereignisse und Größen und nicht diese selbst messe.[33]) Wechsler führt dagegen an, dass „Prognosen nicht mehr als eine Zusammenfassung begründeter, auf dem gegenwärtigen Wissen beruhender zukunftsgerichteter Erwartungen sein können“ und dass das Hinzuziehen von Experten insbesondere bei komplexen Sachverhalten auf jeden Fall legitim sei. Für die eindeutige Beurteilung der „objektiven Validität“ von Vorhersagen wird in der Regel ein Ex-post-Vergleich als notwendig erachtet. Dies ist aber zum Zeitpunkt der Prognoseerstellung nicht möglich. Die „subjektive Validität“ hingegen kann auch zu diesem Zeitpunkt bestimmt werden, indem die logische Konsistenz der erstellten Prognose kontrolliert wird. Auch die eventuell offengelegten Begründungen und Annahmen sind dabei zu überprüfen.[34])

Häder[35]) erarbeitete einen Ansatz, mit dem modellhaft festgestellt werden kann, wie erfolgreich eine Delphi-Studie verlaufen ist. Basis dieser Überlegungen sind zwei unterschiedliche Evaluationskriterien. Die Fehlerverringerung als erstes Kriterium zeigt die Veränderung der Fehlergröße – definiert als Abstand zwischen dem zu schätzenden („wahren“) und dem tatsächlich geschätzten Wert – von Runde zu Runde auf. Dabei gibt die Verringerung des Fehlers mehr oder weniger direkt Auskunft über das Gelingen der Expertenbefragung, weil durch die Delphi-Methode Gruppenprozesse aktiviert werden sollen, in denen zunächst unsichere Urteile schrittweise qualifiziert werden. Die Treffgenauigkeit als zweites Kriterium belegt, ob der wahre Wert innerhalb oder außerhalb der Spannweite der Antworten liegt. Der wahre Wert wird also der Streuung der Schätzungen in der letzten Runde gegenübergestellt. Ein erfolgreicher Verlauf der Delphi-Studie im Sinne der Treffgenauigkeit liegt dann vor, wenn der wahre Wert von der Streuung überdeckt wird. Anhand der Fehlerverringerung kann gezeigt werden, ob sich die Gruppenmeinung in die „richtige Richtung“ bewegt, und die Treffgenauigkeit gibt Auskunft darüber, ob der wahre Wert schließlich getroffen worden ist.

Wenn sich nach der letzten Runde der ursprünglich in der ersten Runde aufgetretene Fehler verringert hat und die Streuung der Expertenschätzungen den wahren Wert überdeckt, kann von einem „Erfolg“ der Delphi-Befragung gesprochen werden. Der Einsatz der Delphi-Methode hat zu keinem verwertbaren Ergebnis geführt, wenn sich der Fehler mit Anzahl der Runden erhöht hat und der wahre Wert

25) Siehe Fußnote 16, S. 730.
26) Siehe Hansmann, K.-W.: „Heuristische Prognoseverfahren“ in WISU – Das Wirtschaftsstudium, Heft 5, 1979, S. 232.
27) Siehe Henze, A.: „Marktforschung“, Stuttgart 1994, S. 22.
28) Siehe Fußnote 27, S. 22.
29) Siehe Sackman, H.: “Delphi Critique: Expert Opinion, Forecasting, and Group Process“, Lexington, Mass., 1975, S. 25.
30) Siehe Wechsler, W.: „Delphi Methode“, Schriftenreihe Wirtschaftswissenschaftliche Forschung und Entwicklung, Band 18, 1978, S. 177.
31) Anhand der informationellen Reproduzierbarkeit kann aufgezeigt werden, inwieweit der zu einem bestimmten Zeitpunkt verfügbare Informationsstand zu übereinstimmenden Vorhersagen führt. Siehe hierzu Fußnote 14, S. 418 f.
32) Siehe Fußnote 27, S. 22.
33) Siehe Fußnote 29, S. 16 und 64 ff.
34) Siehe Fußnote 30, S. 178 ff.
35) Siehe Häder, M: „Zur Evaluation der Delphi-Technik. Eine Ergebnisübersicht“, ZUMA-Arbeitsbericht 96/02, Mannheim 1996.

außerhalb der Spannweite der Antworten liegt. In diesem Fall liegt ein „Misserfolg“ vor. Ist das Kriterium der Fehlerverringerung erfüllt, während sich der wahre Wert außerhalb der Spannweite der Expertenschätzungen befindet, liegt ein (nur) „spezifischer Delphi-Erfolg“ vor. Möglicherweise war die Aufgabenstellung der Befragung in diesem Fall so schwierig, dass die Schätzung noch immer mit einem hohen Maß an Ungenauigkeit verbunden ist, aber dennoch eine erfolgreiche Annäherung an den wahren Wert erreicht werden konnte. Von einem „unspezifischen Delphi-Erfolg“ wird dann gesprochen, wenn sich die Fehlergröße mit den Runden erhöht hat, der wahre Wert aber trotzdem von den Expertenschätzungen überdeckt wird. Hier wäre die für die Delphi-Methode typische Wiederholung der Befragung nicht erforderlich gewesen. Der Erfolg der Befragung wird als „unspezifisch“ bezeichnet, weil der eingetretene Informationsgewinn nicht auf die Delphi-Methode zurückgeführt werden kann, sondern allein auf die Wahl der Expertengruppe.

1.4.2 Experimentelle Überprüfung der Delphi-Methode

Um die Legitimation der Delphi-Methode insbesondere als Prognoseinstrument festzustellen, wurde mittels verschiedener Experimente die Leistungsfähigkeit der Delphi-Methode überprüft. Die Beurteilung erfolgte anhand des Vergleichs der durch die Expertengruppe geschätzten Ergebnisse mit den wahren Werten. Die folgenden Beispiele kann man als typisch bezeichnen.

Die RAND-Corporation untersuchte die Konsistenz der Ergebnisse von Delphi-Befragungen, indem sie Studierenden einen Fragenalmanach vorlegte, deren Antworten ihnen nicht bekannt waren, aber auf der Basis des Allgemeinwissens geschätzt werden konnten. Fragen dieser Art unterscheiden sich zwar von solchen über die Zukunft, sie bieten aber die Gelegenheit festzustellen, ob mit Hilfe der Delphi-Methode eine Annäherung an die richtigen Antworten erzielt werden kann. Tatsächlich näherten sich die Schätzwerte von Runde zu Runde dem wahren Wert an. Außerdem stieg die Konvergenz der Meinungen mit der Anzahl der Befragungsrunden.[36])

Woudenberg berichtet von einem anderen, von der NASA veranstalteten Experiment.[37]) Die Teilnehmer sollten angeben, welche Ausrüstungsgegenstände eine auf dem Mond gestrandete Raumschiffbesatzung zum Überleben benötige, wobei von den NASA-Experten ein Ergebnis als richtig postuliert wurde. Die Delphi-Methode erwies sich im Vergleich zu anderen Ansätzen als diejenige, bei der die beste Schätzung abgegeben wurde. Kritisch an diesem Experiment ist allerdings, ob das als richtig postulierte Ergebnis auch das tatsächlich richtige ist. Eine weitere Möglichkeit zur Evaluation besteht darin, die Delphi-Voraussagen den tatsächlich eingetretenen Ereignissen gegenüberzustellen (Ex-post-Vergleich). Bei einer solchen in Japan durchgeführten Evaluation konnte die Delphi-Methode überzeugen.[38])

1.4.3 Kognitionspsychologische Aspekte

Die Fähigkeit von Experten, richtige Urteile über Sachverhalte zu fällen, die mit Unsicherheit verbunden sind, und die Wiederholung der Befragung bilden die wesentliche Grundlage für die Funktionsweise der Delphi-Methode. Um dies zu fundieren, ist eine kognitionspsychologische Betrachtung des Delphi-Prozesses notwendig.

Den Befragten stehen unterschiedliche Arten von Informationen zur Verfügung. Zum einen sind dies Informationen, die den Befragten wahrscheinlich immer, das heißt unter allen Umständen, in den Sinn kommen. Diese sind im Gedächtnis dauernd verfügbar und ihr Abruf ist damit kontextunabhängig. Zum anderen gibt es Informationen, die nur temporär verfügbar sind. Sie kommen den Befragten nur in einem bestimmten Kontext in den Sinn und sind damit kontextabhängig. Nun werden in einer Befragung nie alle potenziell relevanten Informationen – die zum Urteilszeitpunkt eigentlich verfügbar wären – aus dem Gedächtnis abgerufen, sondern der „Suchprozess“ wird abgebrochen, sobald sich der Befragte an genügend Informationen erinnert hat, um mit hinreichender Sicherheit ein Urteil abgeben zu können. Von welcher Art die abgerufenen Informationen im „Suchprozess“ nun vorrangig sind, hat einen bedeutenden Einfluss auf das Urteil: Dauernd verfügbare, kontextunabhängige Informationen bewirken eine gewisse Stabilität im Urteil, wohingegen temporär verfügbare, kontextabhängige Informationen zu einer gewissen Variabilität im Urteil führen.

Personen, die über besonders viel Wissen im jeweiligen Fachgebiet verfügen, können als Experten im Sinne der Delphi-Methode bezeichnet werden. Die Rückmeldung in der folgenden Runde bewirkt schließlich, dass die Experten den „Suchprozess“ in ihrem Gedächtnis nochmals aufnehmen und nach weiteren, für den jeweiligen Sachverhalt relevanten Informationen fahnden. So bewirkt die zurückgemeldete Gruppenantwort kognitionspsychologisch einen Kontexteffekt und sorgt letztlich in indirekter Weise für eine Verbesserung des abgegebenen Urteils. Die Befragten verfügen über mehr Wissen, als sie zunächst für die Beantwortung einer Frage benutzen. Mittels der Delphi-Methode soll dieses ursprünglich nicht-aktivierte Rest-Wissen aktiviert werden.[39])

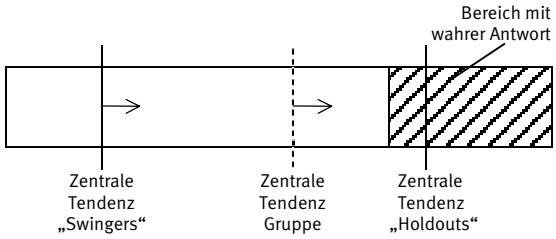
Zur Abschätzung des Urteilsverhaltens von Experten stellt sich die Frage, welcher Mechanismus eine Meinungsänderung in die richtige Richtung bewirkt. So unterscheiden Rowe et al. bei den Teilnehmern zwischen den „Swingers“, die ihre Urteile in einer Folgerunde ändern, da sie in ihrer Haltung unsicher sind und zugleich mit ihren Schätzungen am weitesten vom wahren Wert entfernt liegen, und den „Holdouts“, die am besten geschätzt haben und ihre Antworten nicht ändern, weil sie sich in ihren Urteilen relativ

36) Siehe Fußnote 3, S. 20 ff.
37) Siehe Woudenberg, F.: “An Evaluation of Delphi“ in Technological Forecasting and Social Change, Vol. 40, 1991, S. 131 ff., bes. S. 136 f.
38) Siehe Bundesministerium für Forschung und Technologie (Hrsg.): „Deutscher Delphi-Bericht zur Entwicklung von Wissenschaft und Technik“, Bonn 1993, S. 73 ff.
39) Siehe Häder, M./Häder, S.: „Ergebnisse einer experimentellen Studie zur Delphi-Methode“, ZUMA-Arbeitsbericht 94/05, Mannheim 1994, S. 9.

sicher sind.[40]) Die Gruppenschätzung nähert sich daher mit der Anzahl der Runden immer mehr dem wahren Wert an (siehe Schaubild 1). Die Delphi-Methode ist somit nur funktionsfähig, wenn die Experten über Wissen zur Sicherheit ihres Urteils verfügen. Einige Kritiker sind jedoch der Auffassung, dass die Meinungsänderungen der Experten im Verlauf der Befragung lediglich als Tendenz zum Gruppenmittel und nicht zum vermeintlich wahren Wert zu betrachten sind.

Schaubild 1

Theoretische Änderung des Gruppenurteils mit der Rundenanzahl



Quelle: Eigene Darstellung in Anlehnung an Rowe et al.: “Delphi: A Reevaluation of Research and Theory“ in Technological Forecasting and Social Change, Vol. 39, 1991, S. 238.

Statistisches Bundesamt 2003 - 01 - 0498

Bardecki befasst sich bezüglich der Tendenz zur Meinungsänderung mit der Überlegung, dass die Rückmeldung als externer psychologischer Anker fungiert, der einen bestimmten Einfluss auf die Urteilsbildung haben kann.[41]) Dabei gibt es drei Möglichkeiten:

- der Anker wird nicht beachtet,
- der Anker wird beachtet, indem das Urteil von ihm weg verändert wird, zum Beispiel um die Gesamtmeinung näher an die eigene zu bringen (Kontrasthaltung) oder
- der Anker wird beachtet, indem das Urteil zu ihm hin verändert wird (Assimilationshaltung).

Bardecki vermutet nun, dass je stärker die Einzelschätzung von der Gruppenschätzung abweicht, desto stärker auch der Druck zur Assimilation ausfällt.[42]) Dabei hängt die Abweichung von der Gruppe wiederum von verschiedenen Aspekten ab:

- von der Glaubwürdigkeit der (anderen) Experten,
- von der eigenen Urteilssicherheit,
- von der Gruppengröße und
- von der Bedeutung des Gegenstands.

2 Empirische Überprüfung des erwarteten Delphi-Prozesses

Eine im Laufe des Jahres 2001 am Institut für Agrarpolitik und landwirtschaftliche Marktlehre der Universität Hohenheim durchgeführte Delphi-Befragung hatte zum Ziel, künftige Entwicklungen auf dem Agrartechnikmarkt zu prognostizieren. Im Mittelpunkt standen dabei die Entwicklung in der Landwirtschaft, sowie die Entwicklung der Agrartechniknachfrage und der Anbieterkonzentration. Darüber hinaus sollten noch Tendenzen im Marketing herausgearbeitet werden. Als Fragetypen kamen in einer ersten Befragungsrunde offene Fragen, geschlossene Fragen, bei denen die Teilnehmenden einen konkreten Wert als Schätzung angeben sollten (Identifikationstyp), und geschlossene Fragen, bei denen die Teilnehmenden eine von mehreren vorgegebenen Antwortmöglichkeiten wählen mussten (Selektionstyp), zum Einsatz. In der zweiten Runde wurden die offenen Fragen durch Rangfragen ersetzt und die Befragten hatten die Möglichkeit, Kommentare zu den einzelnen Fragen abzugeben.[43]) Im Folgenden soll nun für die einzelnen Fragekategorien überprüft werden, ob Variationen in der Auswertung (z. B. Gewichtung der Experten, Zerlegung in Gruppen) die Ergebnisse beeinflussen und inwieweit die theoretisch erwarteten Prozesse innerhalb einer Delphi-Befragung auftraten.

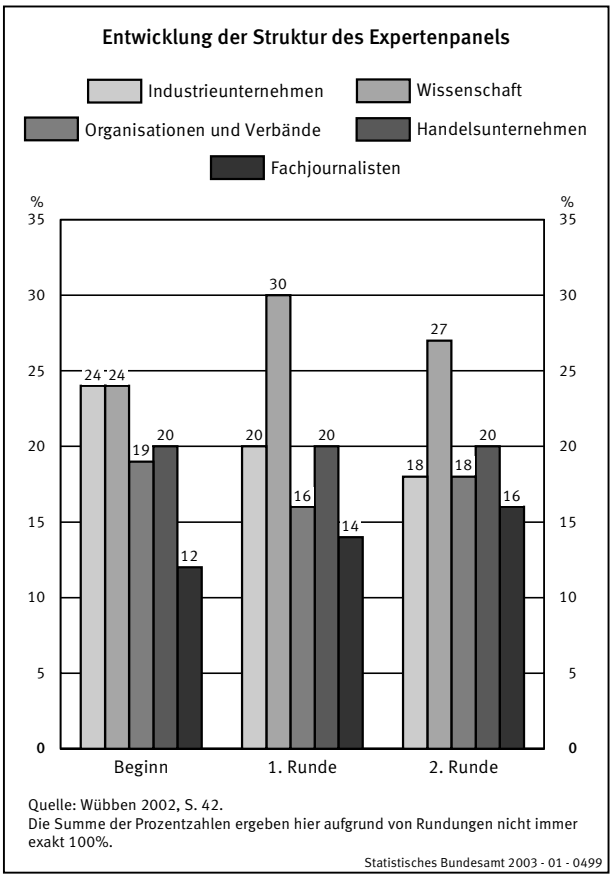
2.1 Anzahl der Befragungsrunden und Panel-Mortalität

Die Durchführung der Befragung beschränkte sich auf zwei Runden. Dafür sprachen zwei Gründe. Zum einen war nach diesen beiden Runden bei den meisten Fragen bereits ein Konsens zwischen den Experten erzielt worden, der als ausreichend angesehen werden konnte. Zum anderen war der Zeitaufwand für eine Befragungsrunde zu groß, als dass eine weitere Fragerunde – die nur noch wenige Fragen beinhaltet hätte – zu rechtfertigen gewesen wäre. Das in Abschnitt 1.3.2 formulierte Kriterium, wonach eine Delphi-Befragung solange durchgeführt wird, bis bei einer minimalen Anzahl von Runden ein annehmbares Maß an Genauigkeit erzielt worden ist, konnte als erfüllt gelten.

In der ersten Runde der Delphi-Befragung wurden insgesamt 83 Fragebogen an die ausgewählten Experten verschickt. Davon wurden 50 Fragebogen beantwortet und für die statistische Analyse verwendet. Die Rücklaufquote lag damit im Bereich vergleichbarer Delphi-Befragungen. Bei der zweiten Runde wurden von 50 Fragebogen 44 beantwortet, was einer Rücklaufquote von 88% entspricht. Im Vergleich zu anderen Delphi-Befragungen lag die Quote hier im überdurchschnittlichen Bereich. Insgesamt wies die Delphi-Befragung eine vergleichsweise geringe Panel-Mortalität auf, was die Rücklaufquote über beide Runden von etwa 53% zeigt. Von den einzelnen Expertengruppen wies die zahlenmäßig kleinste Gruppe der Fachjournalisten mit 70% die höchste Rücklaufquote auf. Die geringste Quote zeigte sich mit 40% bei den Experten aus den Industrieunternehmen.

40) Siehe Rowe, G. et al.: “Delphi: A Reevaluation of Research and Theory” in Technological Forecasting and Social Change, Vol. 39, 1991, S. 237 f.
41) Siehe Bardecki, M.: “Participants’ Response to the Delphi Method” in Technological Forecasting and Social Change, Vol. 25, 1991, S. 282 f.
42) Siehe Fußnote 41, S. 283.
43) Zu den Ergebnissen der Studie siehe Vorgrimler, D./Wübben, D.: „Prognose der Entwicklung des Agrartechnikmarktes – Eine Expertenbefragung nach der Delphi-Methode“ in Hohenheimer Agrarökonomische Arbeitsberichte, Nr. 7, 2001. Internet: *http://www.uni-hohenheim.de/~apo420b/berichte/wpaper.html.*

Schaubild 2



Quelle: Wübben 2002, S. 42.
Die Summe der Prozentzahlen ergeben hier aufgrund von Rundungen nicht immer exakt 100%.

Statistisches Bundesamt 2003 - 01 - 0499

Durch die unterschiedlichen Rücklaufquoten bei den Expertengruppen ändert sich im Laufe der Befragungsrunden auch die Struktur des Expertenpanels. Schaubild 2 zeigt, wie sich die Verteilung der Expertengruppen über den Zeitraum der Befragung entwickelt hat. Aufgrund der unterschiedlichen Rücklaufquoten steigt der Anteil der Gruppe „Fachjournalisten" ausgehend von der Struktur des Anfangs angeschriebenen Panels an, während der Anteil der Expertengruppe „Industrieunternehmen" absinkt.

2.2 Der Einfluss der Gruppen auf die Ergebnisse

Da die Teilnehmer Vertreter fünf unterschiedlicher Gruppen sind, bietet sich eine gruppenspezifische Auswertung der Ergebnisse an. Die Analyse beschränkt sich dabei auf die Fragen des Identifikationstyps (Schätzungsfragen). Tabelle 1 zeigt für sechs ausgewählte Fragen das Gesamt- und die Gruppenergebnisse jeweils absolut und relativ an. Der Vergleich erfolgt anhand der ungewichteten Ergebnisse, um die Effekte auszuschalten, die aus der Gewichtung nach der Selbsteinschätzung der Fachkompetenz resultieren (siehe Abschnitt 2.3).

Anhand des Kruskal-Wallis-Tests[44]) kann untersucht werden, ob sich die jeweiligen Gruppenergebnisse grundsätzlich unterscheiden. Getestet wird die Nullhypothese, dass die mittleren Rangzahlen in den einzelnen Expertengruppen gleich sind, nachdem eine gemeinsame Rangordnung der rangtransformierten Werte der verschiedenen Gruppen erstellt wurde. Die Irrtumswahrscheinlichkeiten bezüglich der Ablehnung dieser Nullhypothese sind zu jeder Frage in der Tabelle 1 dargestellt. Auf dem 5%-Niveau ist signifikant, dass die Gruppenergebnisse bei den Fragen 3 und 6 sowie bei der Frage 4 für 2005 unterschiedlich sind. Bezüglich dieser Fragen entstammen die fünf Expertengruppen mit entsprechender Wahrscheinlichkeit (95%) nicht der gleichen Grundgesamtheit. Auf dem 10%-Niveau sind die

Tabelle 1: Gruppenergebnisse der Schätzungsfragen [absolut und relativ[1])] und Signifikanzniveaus der Unterschiedlichkeit der Urteile/Einschätzungen der Expertengruppen

Schätzfrage	Jahr	Einheit	Insgesamt	Industrie	Wissenschaft	Verbände	Handel	Presse	α [2])
1 Durchschnittliche Betriebsgröße (ha LN)	2005	absolut	57	54,4	55,1	58,3	61,7	55,4	0,250
		relativ	100	95,5	96,7	102,3	108,3	97,3	
	2010	absolut	78,4	83	73	78,3	80	79,9	0,613
		relativ	100	105,8	93,1	99,8	102	101,8	
2 Durchschnittliche Motorleistung (kW) je neu zugelassenem Traktor	2005	absolut	81,1	81,4	78,6	81,7	83,4	81	0,435
		relativ	100	100,3	97	100,8	102,9	99,9	
	2010	absolut	95,1	97,6	88,5	99,3	96,4	96,7	0,078
		relativ	100	102,7	93	104,4	101,4	101,7	
3 Traktorneuzulassungen Deutschland	2005	absolut	21 900	21 100	22 700	22 100	21 800	21 400	0,037
		relativ	100	96,3	103,7	100,7	99,5	97,5	
	2010	absolut	19 700	19 200	20 600	20 100	19 900	18 000	0,030
		relativ	100	97,3	104,7	101,8	100,7	91,1	
4 Verkaufte Mähdrescher Deutschland	2005	absolut	2 080	2 030	2 140	2 060	2 020	2 150	0,016
		relativ	100	97,5	102,7	98,7	97,1	103,2	
	2010	absolut	1 910	1 850	1 980	1 860	1 880	1 920	0,136
		relativ	100	97	103,7	97,8	98,7	101	
5 Gesamtumsatz deutsche Agrartechnikindustrie (Mill. DM)	2005	absolut	6 540	6 610	6 670	6 460	6 410	6 450	0,054
		relativ	100	101,1	102	98,9	98	98,6	
	2010	absolut	6 480	6 570	6 660	6 390	6 320	6 360	0,054
		relativ	100	101,3	102,8	98,5	97,5	98	
6 Umsatzvolumen weltweit (Mrd. US-Dollar)	2005	absolut	54,1	52,7	54	58,1	52,7	52,9	0,000
		relativ	100	97,4	99,8	107,5	97,3	97,8	
	2010	absolut	58	56,3	58,1	61,9	56,1	57,4	0,018
		relativ	100	97,1	100,1	106,8	96,7	99	

1) Bei der relativen Betrachtung ist das jeweilige Gesamtergebnis = 100 gesetzt. – 2) α = Irrtumswahrscheinlichkeiten.

44) Zum Testverfahren siehe Bortz, J. et al.: „Verteilungsfreie Methoden in der Biostatistik", Berlin 2000, S. 222 ff.

Unterschiede für Frage 5 und für das Jahr 2010 für die Frage 2 signifikant. Das bedeutet, dass mindestens eine der fünf Gruppen in ihrer Einschätzung eine andere zentrale Tendenz aufweist als mindestens eine andere. Bei den restlichen Fragen muss auf dem 10%-Signifikanzniveau die Nullhypothese beibehalten werden.

Welche Gruppen sich in ihren Urteilen konkret unterscheiden, kann mit dem ebenfalls nichtparametrischen Mann-Whitney-Test (U-Test)[45]) untersucht werden, indem die Verteilung der rangtransformierten Antworten zwischen zwei Gruppen verglichen werden. Getestet wird die Nullhypothese, dass sich jeweils beide Gruppen nicht unterscheiden, also einer gleichen Grundgesamtheit angehören. Statt eines paarweisen Vergleichs der Gruppen kann mit dem Mann-Whitney-Test auch untersucht werden, ob eine Gruppe jeweils gegenüber dem Rest der Teilnehmer eine andere Tendenz aufweist. Wenn der Test einseitig erfolgt, kann festgestellt werden, ob der Richtungsunterschied der Tendenzen signifikant ist. Der Tabelle 2 sind zu jeder Schätzungsfrage die Signifikanzniveaus der Unterschiede zwischen den jeweiligen Expertengruppen und dem Rest der Teilnehmer zu entnehmen. Es fällt auf, dass besonders die Wissenschaftler in ihren Urteilen oft signifikant von den restlichen Gruppen abweichen. So schätzten sie die Werte bei den Fragen 3 bis 5 größtenteils auf dem Signifikanzniveau von 5% höher – also optimistischer – ein. Bei den Fragen 1 (nur für 2010) und 2 werden von den Wissenschaftlern niedrigere Werte erwartet.

2.3 Der Einfluss der Gewichtung auf das Ergebnis

Durch die Gewichtung wird die relative Wichtigkeit der Befragten geändert. Die Teilnehmenden wurden gebeten, sich zu jeder Frage in eine von drei „Kompetenzgruppen" einzuordnen. In die gewichteten Ergebnisse sind die Antworten, bei denen sich die Experten mit „Fachwissen hoch" einstuften, dreifach eingegangen. Antworten, bei denen sich die Fachleute mit „Fachwissen mittel" einschätzten, sind zweifach in die Berechnung eingeflossen, während Angaben, bei denen die Experten sich nur mit „Fachwissen gering" einstuften, einfach berücksichtigt wurden. Die Veröffentlichung der Ergebnisse erfolgte auf Basis dieser Gewichtung.

Eine Datengewichtung wird in der Regel angewandt, um eine Stichprobe repräsentativ für eine Grundgesamtheit zu machen. Im Falle der vorliegenden Expertenbefragung geht es im Gegensatz dazu nicht um die Repräsentanz einer Grundgesamtheit, sondern ausschließlich darum, ein zuverlässigeres Ergebnis zu erzielen. Das geschieht dadurch, dass mutmaßlich kompetenteren Expertenmeinungen ein höheres Gewicht zugemessen wird.

Die Gewichtung birgt einige Probleme. Die statistischen Eigenschaften von klassischen Tests gehen durch die Gewichtung verloren. So widerspricht die Gewichtung der Stichprobenethik, denn mit einer „Gewichtung lässt sich jedes Ergebnis produzieren"[46]). Daher ist bei der Anwendung von statistischen Tests stets vom ungewichteten Datenmaterial auszugehen. Darüber hinaus ist die Selbsteinschätzung der Experten subjektiven Einflüssen unterworfen. In dem nun folgenden Abschnitt wird gezeigt, inwieweit unterschiedliche Gewichtungen Einfluss auf die Ergebnisse haben.

Tabelle 3 stellt für die Schätzungsfragen das ungewichtete und das 3-2-1-gewichtete[47]) Gesamtergebnis sowie die Gruppenergebnisse der jeweiligen „Kompetenzgruppen" dar. Zusätzlich sind für jede der drei „Kompetenzgruppen" die Anteile der Antworten am Gesamtergebnis aufgeführt.

Zwischen dem 3-2-1-gewichteten und dem ungewichteten (1-1-1-) Ergebnis treten kaum Unterschiede auf. Der

Tabelle 2: Signifikanzniveaus der Unterschiede zwischen den jeweiligen Expertengruppen und dem Rest der Teilnehmer

Schätzungsfrage	Jahr	Industrie	Wissenschaft	Verbände	Handel	Presse
1 Durchschnittliche Betriebsgröße (ha LN)	2005	0,176	0,154	0,366	**0,016**	0,303
	2010	0,343	0,061	0,486	0,173	0,373
2 Durchschnittliche Motorleistung (kW) je neu zugelassenem Traktor	2005	0,448	**0,037**	0,309	0,131	0,303
	2010	0,312	**0,003**	0,072	0,383	0,172
3 Traktorneuzulassungen Deutschland	2005	**0,043**	**0,010**	0,466	0,396	0,202
	2010	0,136	**0,017**	0,453	0,379	**0,041**
4 Verkaufte Mähdrescher Deutschland	2005	0,156	0,057	0,297	0,071	0,097
	2010	0,132	**0,030**	0,160	0,289	0,310
5 Gesamtumsatz deutsche Agrartechnikindustrie (Mill. DM)	2005	0,161	**0,039**	0,221	**0,040**	0,287
	2010	0,159	**0,038**	0,208	0,072	0,198
6 Umsatzvolumen weltweit (Mrd. US-Dollar)	2005	0,107	0,141	**0,011**	0,208	0,051
	2010	0,150	0,118	0,107	0,129	0,298

Jeweilige Gruppe schätzt die Werte auf dem Signifikanzniveau von 10% (fett gedruckt: 5%) niedriger ein.

Jeweilige Gruppe schätzt die Werte auf dem Signifikanzniveau von 10% (fett gedruckt: 5%) höher ein.

Quelle: Wübben 2002, S. 57.

45) Siehe Fußnote 44, S. 200 ff.

46) Gabler, S. et al. (Hrsg.): „Gewichtung in der Umfragepraxis", Opladen 1994, S. 2.

47) Im Folgenden wird die dreifach-, zweifach- und einfach-Gewichtung als 3-2-1-Gewichtung bezeichnet. Dies gilt analog auch für andere „Gewichtungen" (z. B. liegt bei ungewichteter Betrachtung eine 1-1-1-Gewichtung und bei ausschließlicher Berücksichtigung der Gruppe „Fachwissen hoch" eine 1-0-0-Gewichtung vor).

Tabelle 3: Ungewichtetes und 3-2-1-gewichtetes Gesamtergebnis, Einzelergebnisse der „Kompetenzgruppen“ und ihr jeweiliger Anteil am Gesamtergebnis

Schätzungsfrage	Jahr	Einheit	Gesamtergebnis		Gruppenergebnis			Anteil der Antworten bei ungewichtetem Gesamtergebnis in %			Anteil der Antworten bei gewichtetem Gesamtergebnis in %		
					Fachwissen			Fachwissen			Fachwissen		
			ungewichtet	gewichtet	hoch	mittel	gering	hoch	mittel	gering	hoch	mittel	gering
1 Durchschnittliche Betriebsgröße (ha LN)	2005	absolut	57,0	57,7	62,5	58	53,1	*4,8*	*69,0*	*26,2*	*8,0*	*77,3*	*14,7*
		relativ	98,8	100	108,3	100,5	92						
	2010	absolut	78,4	79,2	87,5	79,2	74,8						
		relativ	99	100	110,5	100	94,4						
2 Durchschnittliche Motorleistung (kW) je neu zugelassenem Traktor	2005	absolut	81,1	81,2	79,8	82,3	78,4	*21,4*	*61,9*	*16,7*	*31,4*	*60,5*	*8,1*
		relativ	99,9	100	98,3	101,4	96,6						
	2010	absolut	95,1	94,8	90,3	97,4	92,6						
		relativ	100,3	100	95,3	102,7	97,7						
3 Traktor-neuzulassungen Deutschland	2005	absolut	21 900	21 900	21 400	22 000	22 000	*11,9*	*47,6*	*40,5*	*20,8*	*55,6*	*23,6*
		relativ	100	100	97,7	100,5	100,5						
	2010	absolut	19 700	19 700	19 300	19 900	19 600						
		relativ	100	100	98	101	99,5						
4 Verkaufte Mähdrescher Deutschland	2005	absolut	2 080	2 080	2 080	2 100	2 080	*18,6*	*37,2*	*44,2*	*32,0*	*42,7*	*25,3*
		relativ	100	100	100	101	100						
	2010	absolut	1 910	1 900	1 900	1 910	1 910						
		relativ	100,5	100	100	100,5	100,5						
5 Gesamtumsatz deutsche Agrar-technikindustrie (Mill. DM)	2005	absolut	6 540	6 530	6 450	6 570	6 530	*11,9*	*42,9*	*45,2*	*21,4*	*51,4*	*27,1*
		relativ	100,2	100	98,8	100,6	100						
	2010	absolut	6 480	6 470	6 400	6 480	6 510						
		relativ	100,2	100	98,9	100,2	100,6						
6 Umsatzvolumen weltweit (Mrd. US-Dollar)	2005	absolut	54,1	54,6	54,2	56,3	53	*5,1*	*30,8*	*64,1*	*10,9*	*43,6*	*45,5*
		relativ	99,1	100	99,3	103,1	97,1						
	2010	absolut	58,0	58,7	59,3	60,8	56,6						
		relativ	98,8	100	101	103,6	96,4						

Quelle: Wübben, 2002, S. 62.

maximale Unterschied liegt bei 1,2% [Fragen 1 (2005) und 6 (2010)] und ist damit sehr gering.

Ob sich die Ergebnisse der einzelnen „Kompetenzgruppen“ überhaupt von denen der jeweiligen übrigen Gruppen unterscheiden, kann wiederum mit dem Mann-Whitney-Test (einseitig) untersucht werden. Tabelle 4 zeigt die Irrtumswahrscheinlichkeiten dafür, dass sich eine Gruppe jeweils signifikant von den restlichen Gruppen unterscheidet.

Tabelle 4: Signifikanzniveaus der Unterschiede zwischen den jeweiligen „Kompetenzgruppen“ und dem Rest der Gruppen

Schätzungsfrage	Jahr	„Kompetenzgruppe“		
		Fachwissen		
		hoch	mittel	gering
1 Durchschnittliche Betriebsgröße (ha LN)	2005	0,336	0,174	0,058
	2010	0,388	0,592	0,326
2 Durchschnittliche Motorleistung (kW) je neu zugelassenem Traktor	2005	0,286	0,068	0,145
	2010	0,005	0,002	0,454
3 Traktorneuzulassungen Deutschland	2005	0,366	0,693	0,846
	2010	0,264	0,684	0,747
4 Verkaufte Mähdrescher Deutschland	2005	0,788	0,959	0,794
	2010	0,683	0,929	0,685
5 Gesamtumsatz deutsche Agrartechnik-industrie (Mill. DM)	2005	0,146	0,857	0,443
	2010	0,275	0,683	0,264
6 Umsatzvolumen weltweit (Mrd. US-Dollar)	2005	0,584	0,045	0,029
	2010	0,334	0,047	0,019

Jeweilige Gruppe schätzt die Werte auf dem Signifikanzniveau von 5% niedriger ein.

Jeweilige Gruppe schätzt die Werte auf dem Signifikanzniveau von 5% niedriger ein.

Auffallend ist, dass auf dem 5%-Signifikanzniveau nur bei wenigen Fragen Unterschiede auftreten. Dabei betreffen die Unterschiede nicht in erster Linie die Gruppen mit hoher bzw. geringer Selbsteinschätzung, sondern sind bei allen drei Gruppen gleichermaßen existent. Da keine größeren Unterschiede zwischen den Kompetenzgruppen zu verzeichnen sind, werden die Ergebnisse kaum von einer Gewichtung beeinflusst.

2.4 Die Entwicklung der Streuung der Schätzungen

Der Delphi-Prozess sollte mit einer fortschreitenden Konsensbildung des Gruppenurteils verbunden sein. Anhand von Streuungsmaßen wie zum Beispiel der Standardabweichung kann das Ausmaß eines Konsenses bestimmt werden.

An dieser Stelle soll untersucht werden, ob bei den Schätzungsfragen im Verlauf beider Runden der Delphi-Befragung ein zunehmender Konsens festgestellt werden kann. Weil zu jeder dieser Fragen zwei Streuungen (jeweils eine für jede Befragungsrunde) vorliegen, die Unterschiede im Mittelwert und in der Grundgesamtheit aufweisen, wird als Maßzahl für den Konsens der Variationskoeffizient V verwendet. Hierbei handelt es sich um eine maßstabsunabhängige Standardabweichung, die einen Vergleich der Streuung der ersten Runde mit der der zweiten Runde erlaubt.

Tabelle 5 zeigt die Entwicklung der Variationskoeffizienten bei den Schätzungsfragen, wobei sich die Werte nur auf die Teilnehmer beider Runden beziehen.[48]) Grundsätzlich ist dabei festzustellen, dass in jedem Fall der Variationskoeffizient abnimmt. Es kann also bei jeder dieser Fragen von einer Zunahme des Konsenses im Verlauf der Delphi-Befragung gesprochen werden. Bei den Fragen 1 und 6 (jeweils für 2010) nimmt der Variationskoeffizient sogar um mehr als 50% ab, bei der Frage 2 (für 2005) liegt die Abnahme lediglich bei 10%, wobei hier auch nur eine sehr geringe anfängliche Streuung vorliegt.

Tabelle 5: Variationskoeffizienten der Schätzungsfragen und ihre Veränderung in der zweiten Runde

Schätzungsfrage	Jahr	Variationskoeffizient		Veränderung zweite Runde gegenüber erster Runde	
		erste Runde	zweite Runde	absolut	%
1 Durchschnittliche Betriebsgröße (ha LN)	2005	0,219	0,141	– 0,078	*– 35,6*
	2010	0,368	0,181	– 0,187	*– 50,8*
2 Durchschnittliche Motorleistung (kW) je neu zugelassenem Traktor	2005	0,080	0,072	– 0,008	*– 10,0*
	2010	0,159	0,098	– 0,061	*– 38,4*
3 Traktorneuzulassungen Deutschland	2005	0,090	0,068	– 0,022	*– 24,4*
	2010	0,145	0,100	– 0,045	*– 31,0*
4 Verkaufte Mähdrescher Deutschland	2005	0,117	0,081	– 0,036	*– 30,8*
	2010	0,158	0,091	– 0,067	*– 42,4*
5 Gesamtumsatz deutsche Agrartechnikindustrie (Mill. DM)	2005	0,055	0,044	– 0,011	*– 20,0*
	2010	0,096	0,064	– 0,032	*– 33,3*
6 Umsatzvolumen weltweit (Mrd. US-Dollar)	2005	0,101	0,081	– 0,020	*– 19,8*
	2010	0,196	0,097	– 0,099	*– 50,5*

Quelle: Wübben, 2002, S. 68.

Des Weiteren fällt auf, dass die Variationskoeffizienten bei den Fragen, die sich auf das Jahr 2010 beziehen, jeweils höher sind als diejenigen, die sich auf das Jahr 2005 beziehen, was nach dem Mann-Whitney-Test (einseitig) mit einer Irrtumswahrscheinlichkeit von unter 1% signifikant ist. Dieser Unterschied ist plausibel, weil bei einer Prognose für einen entfernteren Zeitpunkt die Unsicherheit höher und damit auch die Streuung der Einzelurteile größer ist. Die Experten sind allerdings bei den Fragen für das Jahr 2010 aufgrund der Unsicherheit konsensfreudiger. Die prozentuale Abnahme des Variationskoeffizienten ist bei diesen Fragen auf dem 1%-Signifikanzniveau höher als bei den Fragen für das Jahr 2005.

2.5 Die Entwicklung der Rangordnungen

Bei den Rangfragen kann überprüft werden, ob sich die Teilnehmer der zweiten Runde bei der Vergabe der Rangordnungen von der vorgegebenen Rangordnung aus der ersten Runde haben beeinflussen lassen. Inwieweit sich die zurückgemeldete Rangordnung aus der ersten Runde und die resultierende aus der zweiten unterscheiden, lässt sich mittels Kendalls Korrelationskoeffizienten τ bewerten. Gilt $\tau = 1$, stimmen die Rangordnungen überein, bei $\tau = -1$ sind die Rangordnungen gegensätzlich. Ob davon auszugehen ist, dass zwei Rangordnungen als übereinstimmend anzusehen sind, kann mit Kendalls τ-Test untersucht werden.[49]) Tabelle 6 zeigt für jede der 17 in der Studie enthaltenen Rangfragen die Korrelationskoeffizienten τ (nach Kendall) und die Anzahl N der zu ordnenden Objekte (Kategorien). Zusätzlich ist vermerkt, ob die Rangordnungen auf dem 1%-Niveau signifikant nicht unterschiedlich sind.

Tabelle 6: Vergleich zwischen den Rangordnungen beider Befragungsrunden

Nr. der Rangfrage	τ	N	Signifikanz
1	0,93	8	*
2	0,69	10	*
3	0,82	10	*
4	0,67	9	*
5	0,78	10	*
6	0,82	10	*
7	1,00	9	*
8	0,83	9	*
9	1,00	7	*
10	1,00	8	*
11	0,93	8	*
12	0,78	10	*
13	0,60	6	
14	0,60	6	
15	0,71	8	*
16	0,47	6	
17	0,71	8	*

* Auf dem 1%-Niveau signifikant, dass beide Rangordnungen nicht unterschiedlich sind.

Quelle: Wübben, 2002, S. 69.

Beim überwiegenden Teil der Rangfragen kann davon ausgegangen werden, dass sich die Rangordnungen nicht geändert haben. Nur bei den Fragen 13, 14 und 16 kann bei einer Irrtumswahrscheinlichkeit von 1% nicht nachgewiesen werden, dass sich die Rangordnung der zweiten Runde von der der ersten nicht unterscheidet. Diese Fragen weisen mit 6 die geringste Anzahl zu ordnender Kategorien auf. Da die Urteilsfähigkeit (Diskriminanzfähigkeit) der Befragten mit der Anzahl der zu ordnenden Objekte abnimmt[50]), fiel es den Teilnehmenden bei diesen Fragen vermutlich leichter, eine von der ersten Rangordnung „unabhängigere" Reihung vorzunehmen. Bei höherer Anzahl der Kategorien orientierten sich die Experten möglicherweise mehr an der Rangordnung der ersten Runde. Dennoch kann zwischen τ und N keine (lineare) Korrelation (nach Pearson) festgestellt werden ($r = 0{,}34$).

2.6 Die Entwicklung des Urteilsverhalten

Es stellt sich die Frage, inwieweit sich die Teilnehmenden in der zweiten Runde von den Ergebnissen der ersten Runde haben beeinflussen lassen. Grundsätzlich können bei der Delphi-Methode drei Typen von Urteilen unterschieden werden:

– Urteile, die beibehalten werden (stabile Urteile),

48) Werden auch die Experten, die nur an der ersten Runde teilgenommen haben, in die Betrachtung eingeschlossen, könnte dadurch eventuell ein Delphi-Prozess vorgetäuscht werden, der möglicherweise in Wirklichkeit nicht existiert. Das wäre der Fall, wenn mehrere Experten mit einer weit vom Mittelwert abweichenden Schätzung nach der ersten Runde ausfallen würden.

49) Siehe Fußnote 44, S. 422 ff. Die Prüfgröße S bei Kendalls τ-Test ergibt sich aus der Differenz zwischen der Proversionsanzahl (Anzahl der Rangüberschreitungen) und der Inversionsanzahl (Anzahl der Rangunterschreitungen). Die Signifikanz hängt zum einen von S und zum anderen von der Anzahl der zu ordnenden Objekte ab.

50) Siehe Fußnote 6, S. 155.

– Urteile, die sich dem Gruppenurteil annähern (Assimilations-Urteile) und

– Urteile, die sich vom Gruppenurteil in die eigene Richtung entfernen (Kontrast-Urteile).

Werden alle Einzelurteile der Schätzungsfragen aggregiert, erhält man die in Schaubild 3 dargestellten Anteile der Urteilstypen. Es ist ersichtlich, dass außerdem ein vierter Typ nicht-erklärbarer Urteile vorliegt. Hierbei handelt es sich um Antworten, die sich zwar in die Richtung des Gruppenurteils der ersten Runde bewegt haben, dieses jedoch weit übertreffen, sodass sich der Abstand vom Gruppenurteil sogar erhöht hat. Während die drei anderen Urteilstypen auf das Vorhandensein verfügbarer kontextabhängiger bzw. -unabhängiger Informationen zurückzuführen sind, ist nicht klar, wie Antworten dieses vierten Urteilstyps zustande kommen. Es ist daher fraglich, ob diese (Einzel-) Urteile die Prognoseergebnisse verbessern. Der Anteil dieses Urteilstyps beträgt rund 13%. Den größten Anteil mit 55% machen die Assimilations-Urteile aus, was für die Delphi-Methode typisch ist. Rund 25% aller Urteile der Schätzungsfragen sind stabil geblieben und rund 7% weisen eine Kontrasthaltung auf.

Schaubild 3

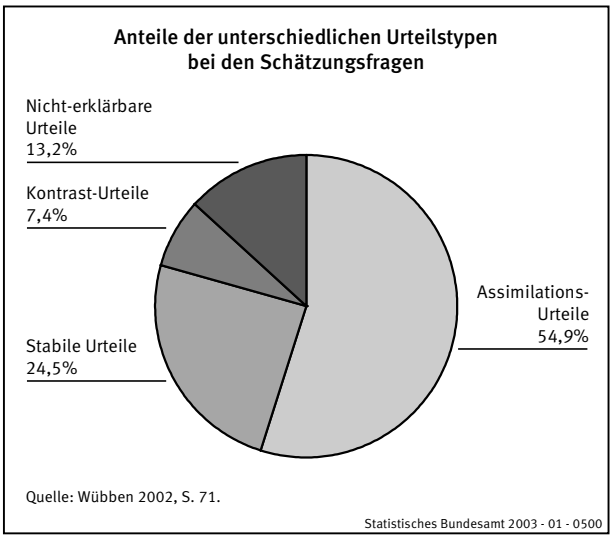


Schaubild 4

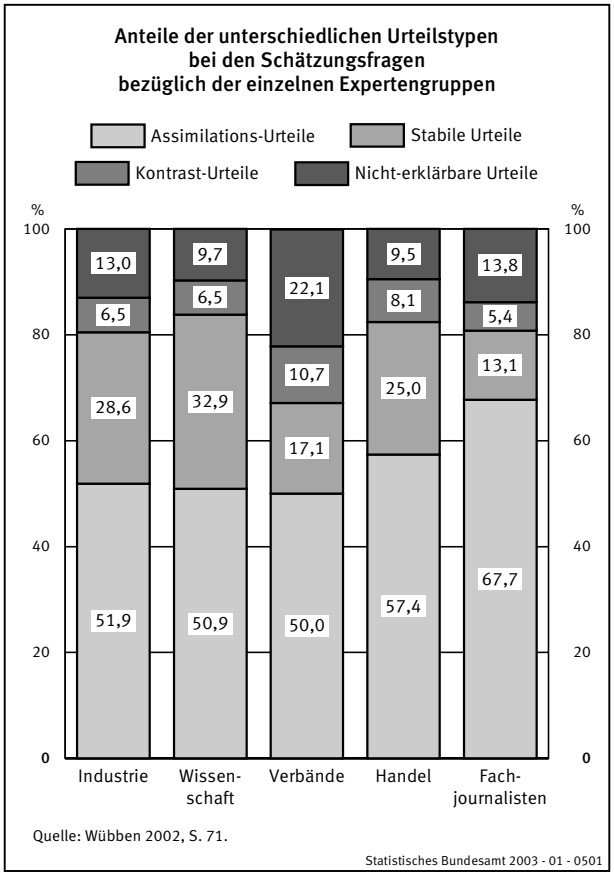


Schaubild 4 zeigt das Urteilsverhalten der einzelnen Expertengruppen in der zweiten Runde. Daraus geht hervor, dass zum einen die Experten aus der Gruppe „Fachjournalisten“ im Urteilsverhalten am „anpassungsfreundlichsten“ (67,7% Assimilations-Urteile und nur 13,1% stabile Urteile) und zum anderen die Vertreter aus der Wissenschaft am „beharrlichsten“ (32,9% stabile Urteile) sind. Mit gut 22% bzw. knapp 11% weisen die Teilnehmer aus Verbänden und Organisationen den jeweils größten Anteil von Kontrast- bzw. nicht-erklärbaren Urteilen auf.

3 Fazit

Die im Jahr 2001 durchgeführte Delphi-Studie hatte größtenteils die Eigenschaften, die von solchen Expertenbefragungen theoretisch erwartet werden. Die Experten näherten sich in ihrer Meinung einem Wert an, was sich durch die Reduzierung der Streuung in den Antworten zeigte. Auch bestätigte sich die Vermutung, dass Experten, die aus verschiedenen gesellschaftlichen Gruppen stammen, unterschiedliche Meinungen besitzen können. Daher verdient die Struktur des Expertenpanels besondere Beachtung. Keine Auswirkung hatte allerdings die Gewichtung der Antworten der Experten aufgrund deren Selbsteinschätzung der eigenen Expertise. Dies deutet entweder darauf hin, dass das subjektive Element bei der Selbsteinschätzung eine zu große Rolle gespielt hat, oder, dass es weniger auf das letzte Detailwissen, als vielmehr darauf ankommt, dass ein Experte sich überhaupt als Experte bezeichnen kann. Ein weiterer Aspekt ist, dass sich die Experten durch die Rückmeldung der Ergebnisse der ersten Runde gemeinsam haben beeinflussen lassen. Dadurch hat sich das unterschiedliche Expertenwissen eventuell nivelliert.

Insgesamt ist die Eignung des Instrumentes Delphi-Methode zur Vorhersage von zukünftigen Ereignissen oder einfach nur zur Abschätzung eines bestimmten zukünftigen Bedarfs positiv zu beurteilen. Als mögliches Anwendungsbeispiel im Bereich der amtlichen Statistik könnte man sich vorstellen, bei den Nutzern der amtlichen Statistik, die in diesem Fall die Experten darstellen, den künftigen Bedarf an statistischen Ergebnissen mit Hilfe der Delphi-Methode zu beurteilen.

Auszug aus Wirtschaft und Statistik

Herausgeber: Statistisches Bundesamt, Wiesbaden

Schriftleitung: N. N.
Verantwortlich für den Inhalt:
Brigitte Reimann,
65180 Wiesbaden

- Telefon: +49 (0) 6 11/75 20 86
- E-Mail: wirtschaft-und-statistik@destatis.de

Vertriebspartner: SFG Servicecenter Fachverlage
Part of the Elsevier Group
Postfach 43 43
72774 Reutlingen
Telefon: +49 (0) 70 71/93 53 50
Telefax: +49 (0) 70 71/93 53 35
E-Mail: destatis@s-f-g.com

Erscheinungsfolge: monatlich

Allgemeine Informationen über das Statistische Bundesamt und sein Datenangebot erhalten Sie:

- im Internet: www.destatis.de

oder bei unserem Informationsservice
65180 Wiesbaden

- Telefon: +49 (0) 6 11/75 24 05
- Telefax: +49 (0) 6 11/75 33 30
- www.destatis.de/kontakt