Grundlagenbeitrag: Quantitative Befragungen

Christoph Böhmert und Ferdinand Abacioglu

Zusammenfassung

Die quantitative Befragung ermöglicht eine effiziente Erhebung und Auswertung großer Mengen an Daten und hat somit ihren festen Platz im Methodenkanon vieler empirischer Wissenschaften. Der vorliegende Beitrag setzt sich im Rahmen der Evaluation von Wissenschaftskommunikation mit den Grundlagen dieses Erhebungsverfahrens auseinander und geht dabei zunächst auf wesentliche Merkmale und Klassifizierungsmöglichkeiten der Methode ein. In Hinblick auf Konzeption und Durchführung werden dann zentrale Anforderungen an das entsprechende Messinstrument, sowie die grundlegenden Eigenschaften von Messskalen und Antwortformaten dargelegt. Der Beitrag gibt zudem praktische Hinweise für die Verwendung von Ratingskalen und schließt mit einer Zusammenfassung bewährter Praktiken bei der allgemeinen Fragenkonstruktion.

Häufig werden in der Forschung quantitative und qualitative methodische Ansätze voneinander abgegrenzt (Döring und Bortz 2016, S. 14 ff.). Wenngleich die Unterscheidung der beiden Ansätze eine lange Tradition hat, lassen sich die Ansätze auch in sogenannten Mixed-Methods-Ansätzen im Rahmen von Forschungsprojekten miteinander kombinieren und integrieren, sodass das

C. Böhmert (✉)
IU Internationale Hochschule, Karlsruhe, Deutschland
E-Mail: christoph.boehmert@iu.org

F. Abacioglu
IU Internationale Hochschule, Frankfurt am Main, Deutschland
E-Mail: ferdinand.abacioglu@iu.org

P. Niemann et al. (Hrsg.), *Evaluationsmethoden der Wissenschaftskommunikation,* https://doi.org/10.1007/978-3-658-39582-7_5

„jeweilige Forschungsproblem umfassender bearbeitet werden kann […] und eine bessere Absicherung der Ergebnisse möglich ist“ (Döring und Bortz 2016, S. 17). Der vorliegende Beitrag beschäftigt sich mit quantitativen Befragungen, es sei aber explizit darauf verwiesen, dass auch im Rahmen von Befragungen quantitative und qualitative Methoden kombiniert werden können. Tatsächlich werden in der wissenschaftskommunikativen Evaluationspraxis sehr häufig quantitative Befragungen (etwa die Bewertung eines Vortrags, des Auftretens des Vortragenden etc. mit einer Schulnote) mit qualitativen Elementen (etwa die offene Frage danach, was am Vortrag besonders gut war und was hätte besser sein können) ergänzt.

Kernmerkmal quantitativer Sozialforschung (im Gegensatz zu qualitativer Sozialforschung) ist, dass numerische Daten, also Messwerte erhoben werden. Solche Messwerte können auf unterschiedliche Arten erhoben werden (Döring und Bortz 2016, S. 321 ff.): In der Evaluationsforschung von Bedeutung sind neben den hier thematisierten Befragungen auch Beobachtungen (z. B. wie sich die Zuschauer im Rahmen eines Science Slams verhalten, zu Beobachtungen allgemein siehe auch Weiß in diesem Band) und Tests (z. B. Wissenstests im Anschluss an auf Wissensvermittlung ausgelegte Formate der Wissenschaftskommunikation, siehe auch Wirth und Fleischer in diesem Band).

Bei quantitativen Befragungen wird in der Regel ein möglichst hoher Grad an *Standardisierung* der Befragung angestrebt (Reinecke 2019, S. 717). Diese bezieht sich auf den Grad der Festlegung

- des Fragetextes,
- der Antwortkategorien sowie
- der Reihenfolge der Fragen.

1 Taxonomie quantitativer Befragungen

Quantitative Befragungen können hinsichtlich verschiedener Aspekte klassifiziert werden. Diese Klassifikationen werden hier zunächst eingeführt und weiter unten in diesem Abschnitt an Beispielen aus der Evaluation von Wissenschaftskommunikation verdeutlicht.

Für die Durchführung einer Befragung lassen sich klassischerweise die drei Befragungsarten persönlich-mündlich, telefonisch und schriftlich ausmachen (Fuchs 2019). Neben diesen Arten einer Befragung (auch Modi genannt) sind allerdings insbesondere auch Onlinebefragungen und mobile Befragungen zu erwähnen. Zur weiteren Differenzierung von Befragungen werden in der Literatur

daher häufig die Dimensionen *Administrationsform, Kommunikationskanal* und *Befragungstechnologie* herangezogen (Faulbaum et al. 2009; Reinecke 2019). So identifiziert die Administrationsform einer Befragung, von wem die Fragen (vor-) gelesen beziehungsweise die Antworten registriert werden. Dies kann entweder durch die befragte Person selbst geschehen oder durch eine interviewende Person. Der primäre Kommunikationskanal kann visuell oder auditiv sein, wobei auch Mischformen möglich sind. Als dritte Dimension unterscheidet die Befragungstechnologie zunächst schlicht zwischen der reinen An- und Abwesenheit einer technologischen Unterstützung (Faulbaum et al. 2009). So kann beispielsweise eine telefonische Befragung ohne weitere Unterstützung durchgeführt werden oder durch den Einsatz entsprechender Software aufseiten der Interviewenden computergestützt sein. In Hinblick auf unterschiedliche Kommunikationstechnologien und Endgeräte – welche grundlegenden Einfluss auf die Art einer Befragung nehmen können – ist jedoch anzumerken, dass die Befragungstechnologie nicht als reine Unterstützung der bereits angesprochenen klassischen Modi verstanden werden soll: Je nach Befragungstechnologie lassen sich zusammen mit Administrationsform und Kommunikationskanal spezifische Befragungsarten abbilden.

Eine weitere Klassifizierungsmöglichkeit stellen die Befragten selbst dar: Wer wird überhaupt befragt? Bei der Evaluation von Wissenschaftskommunikation kann hier nach den Rezipient:innen, den Kommunikator:innen sowie weiteren Stakeholder:innen (z. B. Auftraggeber:innen) unterschieden werden. Innerhalb der Gruppe der Rezipient:innen kann nochmals zwischen Rezipient:innen, auf die die Kommunikation abzielt (Zielgruppe), und Rezipient:innen, auf die diese nicht explizit abzielt (sonstige Rezipient:innen), unterschieden werden. Der Rezipient:innenbegriff umfasst hier sowohl aktive, mit den Kommunikator:innen in Dialog stehende als auch passive Rezipient:innen. Abb. 1 visualisiert diese Taxonomie quantitativer Befragungen im Rahmen der Evaluation von Wissenschaftskommunikation.

Darüber hinaus können Befragungen auch hinsichtlich des Befragungssettings (Befragung von Einzelpersonen vs. einer Gruppe von Personen) sowie hinsichtlich des Befragungsgegenstands unterschieden werden (Raithel 2008).

Bei Befragungen in der Evaluation ebenfalls von großer Relevanz sind generelle Aspekte des Forschungsdesigns. Diese spielen auch bei anderen Evaluationsmethoden eine große Rolle. Zu nennen ist hier insbesondere die Planung des Untersuchungszeitpunkts bzw. auch mehrerer Untersuchungszeitpunkte (Döring und Bortz 2016).

Betrachten wir zur Veranschaulichung zwei Beispiele, in denen Wissenschaftskommunikation evaluiert wird:

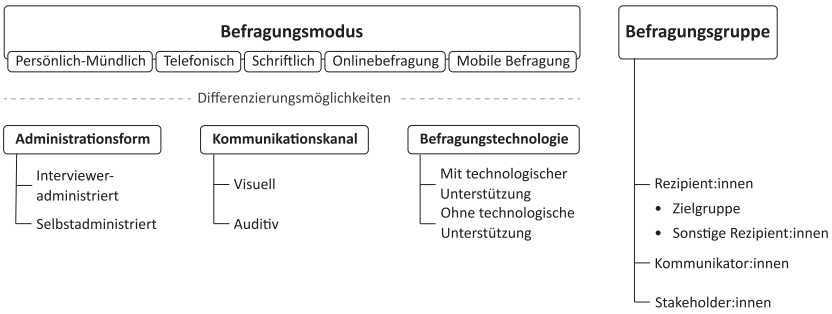

Abb. 1 Klassifizierungsmodell quantitativer Befragungen im Rahmen der Evaluation von Wissenschaftskommunikation. Befragungsmodus in Anlehnung an Faulbaum et al. (2009) und Reinecke (2019)

1. Bewertung von Science Slam-Beiträgen durch das Publikum
2. Bewertung der Verständlichkeit von wissenschaftsjournalistischen Texten

Beispiel 1: Am Ende eines Science Slams bewertet das Publikum alle gehaltenen Vorträge. Dazu findet sich das Publikum in mehreren Gruppen zusammen. In diesen Gruppen werden die Vorträge diskutiert und anschließend Punkte für den Vortrag (oder mehrere Punktwerte für unterschiedliche Aspekte des Vortrags) vergeben. Die jeweiligen Punkte werden den Organisator:innen der Veranstaltung anschließend durch das Hochhalten von Bewertungstafeln (mit den Nummern 1 bis 10) mitgeteilt. Wie lässt sich diese Form der Bewertung klassifizieren? Es handelt sich offenbar um eine standardisierte Befragung, da die Anweisungen, welche die Moderatorin über das Mikrofon gibt, für alle bewertenden Gruppen gleich sind, die Antwortkategorien (Tafeln mit Nummern von 1 bis 10) vorgegeben sind, und die Abfrage für sämtliche Gruppen in derselben Reihenfolge stattfindet (zunächst Bewertung von Vortrag 1, dann von Vortrag 2 etc.). Die Befragung ist durch eine interviewende Person (die Moderatorin) administriert. Der Kommunikationskanal bezieht sich nach Faulbaum et al. (2009) auf den Kanal, über den Befragte ihren Input erhalten. In unserem Science Slam Beispiel geschieht dieser Input im Wesentlichen über den auditiven Kanal (die Moderatorin formuliert die Frage mündlich), zum Teil aber auch über den visuellen (z. B. Mimik und Gestik der Moderatorin). Die Befragung geschieht persönlich und es wird die Gruppe der Rezipient:innen befragt.

Beispiel 2: In einer Studie von Böhmert et al. bewerteten Proband:innen (Studierende) wissenschaftsjournalistische Kurznachrichten unter anderem hin-

weitschweifig	+++	++	+	O	+	++	+++	aufs Wesentliche beschränkt

Abb. 2 Beispielitem gemäß der Verständlichkeitsskala von Langer et al. (2019)

sichtlich ihrer Verständlichkeit (Böhmert et al. 2021). Die Kurznachrichten sowie die Fragen dazu erhielten sie im Hörsaal am Ende einer Lehrveranstaltung. Zur Messung der Verständlichkeit wurde eine bestehende Skala von Langer et al. (2019) adaptiert. Die Skala besteht aus insgesamt 20 Items, die vier zugrunde liegende Dimensionen quantifizieren. Ein Beispiel-Item für die Dimension „Kürze/Prägnanz" ist in Abb. 2 gezeigt. Es handelt sich hier ebenfalls um eine standardisierte Befragung, bei der alle Teilnehmenden exakt dieselbe Instruktion, dieselben Fragen und Antwortmöglichkeiten sowie dieselbe Reihenfolge der Fragen erhielten. Die Instruktionen waren in Schriftform im Fragebogen gegeben, daher handelte es sich um eine selbstadministrierte Befragung über einen visuellen Kommunikationskanal. Die genutzte Befragungstechnologie war „persönlich" (ohne technische Hilfsmittel, einsammeln der Fragebogen am Ende der Vorlesung) und befragt wurden ebenfalls die Rezipient:innen.

2 Hauptgütekriterien

Um die interessierenden Merkmale eines Evaluationsgegenstandes zu untersuchen, bedarf es geeigneter Messinstrumente. Dabei kann die Qualität der Instrumente anhand mehrerer Gütekriterien überprüft werden: Im Rahmen der quantitativen Forschung werden darunter Anforderungen verstanden, welche sich im Wesentlichen auf Konzeption und Anwendung der Messinstrumente beziehen. Unterschieden werden die drei Hauptgütekriterien der Objektivität, Reliabilität und Validität. Diese Gütekriterien müssen im Rahmen von Befragungen soweit wie möglich sichergestellt werden. Nur dann kann davon ausgegangen werden, dass eine Befragung Daten von hoher Qualität liefert.

2.1 Objektivität

Ein grundlegender Anspruch an verwendete Messinstrumente ist die Vergleichbarkeit ihrer Ergebnisse: Werden Daten nicht immer auf die gleiche Weise

erhoben, ausgewertet und interpretiert, unterliegt die Messung Störeinflüssen, welche die gewonnen Ergebnisse verzerren können (Moosbrugger und Kelava 2020). Das Kriterium der Objektivität bezeichnet daher die Anforderung, dass äußere Einflüsse konstant gehalten werden müssen. Befragungen werden von unterschiedlichen Testleiter:innen immer gleich durchgeführt und ein gegebener Datensatz wird von ihnen immer auf die gleiche Weise ausgewertet und interpretiert. Hierfür können drei Dimensionen des Kriteriums unterschieden werden, welche sich auf diesen Prozess der Messung beziehen: Durchführungsobjektivität, Auswertungsobjektivität und Interpretationsobjektivität (Moosbrugger und Kelava 2020). Die objektive Durchführung einer Befragung erfordert in erster Linie eine Standardisierung der Fragetexte und der Reihenfolge der Fragen. Für eine objektive Auswertung ist unter anderem die Standardisierung der Antwortkategorien Voraussetzung. Ist auch die Interpretation der numerischen Ergebnisse hinsichtlich des zu evaluierenden Merkmals standardisiert, kann die Interpretationsobjektivität als gegeben angesehen werden.

2.2 Reliabilität

Im Gegensatz zur Objektivität, welche sich auf äußere Einflüsse bezieht, beschreibt Reliabilität eine Eigenschaft des Messinstruments beziehungsweise der Messung selbst. Von einem reliablen, d. h. zuverlässigen Instrument wird erwartet, dass es unter gleichen äußeren Bedingungen (gegebene Objektivität) für denselben, unveränderten Untersuchungsgegenstand immer wieder zu den gleichen Ergebnissen kommt – das bedeutet, dass Ergebnisse grundlegend reproduzierbar sind. Häufig ist eine exakte Wiederholung der Ergebnisse allerdings nicht möglich: In der Praxis unterliegen Messungen gemäß der sogenannten klassischen Testtheorie immer einem gewissen zufälligen Messfehler, welcher dafür sorgt, dass der gemessene Wert nicht exakt mit dem wahren Wert (beispielsweise exakte Meinung einer befragten Person) übereinstimmt (Moosbrugger und Kelava 2020). Je stärker ein Instrument messfehlerbehaftet ist, desto unzuverlässiger und unpräziser kann das Merkmal des Evaluationsgegenstandes nur untersucht werden und entsprechend gering ist die Reliabilität der Befragung. Doch wie kann die Reliabilität quantifiziert werden? Wird sie als Zuverlässigkeit eines Messinstrumentes im Sinne der Reproduzierbarkeit von Ergebnissen verstanden, kann eine Messung schlicht zu einem zweiten Zeitpunkt wiederholt werden. Je stärker die Ergebnisse beider Messungen miteinander zusammenhängen, desto geringer ist der Messfehler und entsprechend hoch die Reliabilität. Dieses als „Retest“ bezeichnete Verfahren beruht allerdings

auf der Annahme, dass der Untersuchungsgegenstand zwischen den zwei Messzeitpunkten unverändert bleibt. Einstellungen und Meinungen von Befragten, die im Zuge einer Evaluation erhoben werden, können sich jedoch über die Zeit verändern. In diesem Fall würde durch einen Retest nicht nur die Reliabilität, sondern auch der Grad besagter Einstellungs- oder Meinungsänderung erfasst. Neben der Retest-Reliabilität existieren mit der Paralleltestreliabilität, Split-Half-Reliabilität und Konsistenzanalyse weitere Verfahren zur Reliabilitätsbestimmung, welche keinem entsprechenden Zeiteffekt unterliegen. Eine Übersicht dieser klassischen Reliabilitätsmaße geben Moosbrugger und Kelava (2020); Cronbachs Alpha (Cronbach 1951) ist dabei eines der am häufigsten genutzten Maße. Der Koeffizient bildet die interne Konsistenz einer verwendeten Skala (eine Gruppe an Fragen, welche zusammen dasselbe Merkmal erfassen sollen) ab. Alpha gibt also darüber Auskunft, wie stark die einzelnen Fragen miteinander zusammenhängen. Wird angenommen, dass jede dieser Fragen einen eigenen „Test" darstellt und all diese Tests dasselbe Merkmal des Evaluationsgegenstandes erfassen, ergibt sich aus einer hohen internen Konsistenz auch eine hohe Reliabilität: Nur wenn der Messfehler der einzelnen als Test angenommenen Fragen gering ist, können sie auch hoch miteinander korrelieren. An dieser Stelle sei noch darauf hingewiesen, dass vor der Reliabilitätsbestimmung mit den oben beschriebenen Verfahren jeweils bestimmte messtheoretische Voraussetzungen überprüft werden müssen (siehe hierzu Gäde et al. 2020).

2.3 Validität

Als wichtigstes Gütekriterium hinsichtlich der praktischen Anwendung von Messinstrumenten kann die Validität (Gültigkeit) angesehen werden (Moosbrugger und Kelava 2020). Validität bedeutet, dass ein Test das misst, was er vorgibt zu messen (und nicht irgendetwas anderes). Objektivität und Reliabilität sind notwendige aber keine hinreichenden Voraussetzungen für die Validität (Moosbrugger und Kelava 2020). Eine quantitative Befragung kann also nur dann das messen, was sie messen soll (Validität), wenn sie unabhängig von äußeren Umständen (objektiv) und ausreichend zuverlässig (reliabel) Ergebnisse liefert. Es ist jedoch auch möglich, dass die Befragung, obwohl sie den Ansprüchen der Objektivität und Reliabilität genügt, in Wirklichkeit etwas anderes misst, als sie soll.

Evaluation von Wissenschaftskommunikation befasst sich mit der Qualität der Kommunikation. Doch was macht die Qualität von Wissenschaftskommunikation in einem bestimmten Kontext aus? Muss die Kommunikation

in erster Linie verständlich sein? Sollte sie zudem unterhaltsam sein? Sollte sie zum Handeln anregen? Diese und weitere Fragen müssen bei der Evaluation zunächst beantwortet werden, d. h. es muss zunächst geklärt werden, was eigentlich das Kommunikationsziel ist, und wie dieses gemessen werden soll. Erst im Anschluss daran können Evaluator:innen sich mit der Frage beschäftigen, wie sie eine quantitative Befragung so gestalten können, dass diese auch tatsächlich die Kommunikation evaluiert.

Bei der Sicherstellung der Validität sollten vier verschiedene Aspekte betrachtet werden (Moosbrugger und Kelava 2020). Ein zentraler Aspekt der Gültigkeit ist die sogenannte *Kriteriumsvalidität*. Diese ist dann gegeben, wenn ein Zusammenhang zwischen den Angaben einer Person in einer Befragung und ihrem Verhalten außerhalb der Befragungssituation besteht. In unserem Science-Slam-Beispiel wäre die Befragung dann kriteriumsvalide, wenn eine hohe Punktzahl des Vortrags mit anschließendem Verhalten einhergeht, wie etwa, dass die Befragten über gut bewertete Vorträge anschließend mehr recherchieren, diskutieren, diese weiterempfehlen etc. – und über schlecht bewertete Vorträge entsprechend nicht.

Ein zweiter, zentraler Aspekt ist die *Konstruktvalidität*. Konstruktvalidität ist dann gegeben, wenn die Struktur der Antworten auf den Fragebogen sowie deren Übereinstimmungen mit und Abgrenzung von Antworten auf anderen Messinstrumenten mit der zugrunde liegenden Theorie konform ist (Moosbrugger und Kelava 2020). Ein Aspekt im Rahmen der Evaluation von Wissenschaftskommunikation kann hier beispielsweise sein, dass man einen Zusammenhang zwischen der generellen Bewertung und der Bewertung der Verständlichkeit erwartet. Ein „perfekter“ Zusammenhang wäre allerdings nicht zu erwarten, da man gemäß Theorie davon ausgehen würde, dass Verständlichkeit nur einen Teilaspekt der generellen Bewertung ausmacht.

Die beiden weiteren Aspekte der Validität, die *Augenscheinvalidität* und die *Inhaltsvalidität,* behandeln im Gegensatz zur Kriteriumsvalidität und Konstruktvalidität nicht die Antworten auf den Fragebogen, sondern den Fragebogen in seiner Gestaltung selbst. Augenscheinvalidität ist dann gegeben, wenn Laien davon ausgehen, dass ein Befragungsinstrument auch tatsächlich das misst, was es vorgibt zu messen. Inhaltsvalidität liegt dann vor, wenn auch Expert:innen zu diesem Schluss kommen.

3 Skalenniveau und Antwortformate

Auf Bitten der Moderatorin des Science Slams hält eine Gruppe von Zuschauer:innen ihre Punktzahl – in diesem Fall die Bestnote zehn – für einen Science Slam zum Thema Astrophysik in die Höhe. Technisch betrachtet findet in diesem Moment eine Messung statt, und zwar die Messung des Merkmals „Einschätzung der Vortragsqualität". Merkmalsträger ist die Gruppe, die das Schild hochhält. Die gemessene Merkmalsausprägung ist in diesem Falle die zehn. Bei Messungen erfolgt ganz allgemein die Zuordnung eines numerischen Relativs (einer Zahl) zu einem empirischen Relativ (einer Merkmalsausprägung, z. B. „besonders positive Bewertung des Vortrags"; Wirtz und Nachtigall 2012). Vom zu messenden Merkmal selbst sowie von der Art der Messung hängt nun ab, welche Aussagen Evaluierende anhand ihrer Daten treffen können und welche Kennwerte (z. B. Mittelwert etc.) sie berechnen können. Man spricht in diesem Zusammenhang von verschiedenen Skalenniveaus. Das niedrigste Skalenniveau stellt die *Nominalskala* dar. Anhand nominalskalierter Messungen lassen sich lediglich Aussagen über Gleichheit oder Verschiedenheit treffen. Ein Beispiel ist die Messung des Merkmals Geschlecht mit dreistufiger Skala (0 = männlich, 1 = weiblich, 2 = anderes Geschlecht). Das nächsthöhere Skalenniveau ist die *Ordinalskala*. Bei ordinalskalierten Messungen spiegelt das numerische Relativ Verschiedenheit und zusätzlich eine Rangordnung im empirischen Relativ wider. Ein Beispiel hierfür wäre eine grobe Messung des Merkmals „Alter des Publikums" mit einer vierstufigen Skala (1 = unter 18, 2 = 18 bis 39, 3 = 40 bis 59, 4 = 60 und älter). Das über der Ordinalskala liegende Skalenniveau ist die *Intervallskala*. Intervallskalierte Messungen bilden im numerischen Relativ Verschiedenheit, Rangordnung und zusätzlich eine Gleichheit der Abstände (Äquidistanz) ab. Gehen wir in unserem Science Slam-Beispiel davon aus, dass der Unterschied im empirischen Relativ zwischen beispielsweise 2 Punkten und 4 Punkten gleich groß ist wie der zwischen 7 Punkten und 9 Punkten, dann gehen wir von einer intervallskalierten Messung aus (Hussy et al. 2013). Intervallskalierte Messung ist eine wichtige Voraussetzung vieler Auswertungsverfahren. Beispielsweise können Mittelwerte nur für mindestens intervallskalierte Merkmale[1] sinnvoll interpretiert werden. Das Skalenniveau ist ein entscheidendes Element, das bei der Entwicklung von quantitativen Befragungsinstrumenten mit bedacht werden muss.

[1] Das höchste Skalenniveau stellt die Verhältnisskala dar, die hier jedoch nicht weiter thematisiert werden soll (z. B. Rasch et al. 2009, S. 12).

Fragen werden mitunter anhand ihres Inhalts als auch ihrer Form unterschieden (Porst 2014). Inhaltlich können Fragen zu Einstellungen, Überzeugungen, Wissen, Verhalten oder Merkmalen einer befragten Person grob unterschieden werden. Die Form einer Frage bezieht sich hingegen darauf, wie eine Antwort darauf gegeben werden kann. Unterschieden werden hier geschlossene und offene Fragen, wobei auch Mischformen (halboffene Fragen) möglich sind, welche die zwei Fragetypen miteinander verbinden (Porst 2014). Geschlossene Fragen geben eine feste Anzahl an möglichen Antworten vor, aus welchen die befragte Person eine einzige (Einfachnennung) oder potenziell mehrere (Mehrfachnennung) auswählen kann. Ein Beispiel für eine geschlossene Frage mit Einfachnennung ist die Abfrage des Geschlechts. Dem gegenüber stehen offene Fragen, welche keine festen Antwortkategorien vorgeben. Im Rahmen quantitativer Forschung könnte hierfür die freie Angabe des Alters in Jahren als Beispiel dienen. Schlussendlich stellen halboffene Fragen eine Erweiterung des geschlossenen Typs dar, indem sie neben vorgegebenen Antwortalternativen zusätzlich die Option einer offenen Antwort ermöglichen. Geschlossene Fragen erlauben durch ihr vorgegebenes Antwortformat eine standardisierte Befragung. Neben der reinen Ein- und Mehrfachnennung von Antwortalternativen kann mit geschlossenen Fragen auch eine Abfrage von Einstellungen und Meinungen vorgenommen werden. In diesem Fall besitzen die Antwortkategorien eine Rangfolge: Die Wertungen von eins bis zehn, welche im Beispiel des Science Slams etwa für die allgemeine Qualität eines Vortrags vergeben werden, bilden einzelne Rangplätze ab und gelten daher als mindestens ordinalskaliert. Für die Auswertung solcher „Ratingskalen“ ist es günstig, wenn sie Intervallskalenniveau aufweisen. Abb. 3 zeigt eine Auswahl der vorgestellten Frageformen.

3.1 Skalenbeschriftung, Polarität und Anzahl der Antwortkategorien

Die erwähnte „Science-Slam-Skala“ von eins bis zehn Punkten beschränkt sich auf verbale Definitionen ihrer Endpunkte: Die Moderatorin erklärt, dass die eins für die schlechteste und die zehn für die beste Vortragsqualität steht. Soll in unserem Beispiel neben der allgemeinen Qualität des Vortrags auch der Unterhaltungswert erfasst werden, könnte ein Punkt das Urteil „sehr langweilig“ abbilden und zehn Punkte entsprechend das Urteil „sehr unterhaltsam“ – die Verbalisierung der Notenpunkte dazwischen bleibt allerdings aus. Eine Alternative zu dieser hauptsächlich numerischen Antwortskala bilden Antwortformate,

A **Wie ist Ihr Geschlecht?**
- ○ Männlich
- ○ Weiblich
- ○ Anderes Geschlecht

B **Bitte geben Sie an, welche Bereiche der Forschung Sie am meisten interessieren.**
Mehrfachnennung möglich.
- ☐ Life Sciences
- ☐ Biologie
- ☐ Chemie
- ☐ Medizin
- ☐ Anderes: _____

C **Der Inhalt des Science-Slam-Vortrags war unterhaltsam.**

Trifft zu	Trifft eher zu	Trifft eher nicht zu	Trifft nicht zu
○	○	○	○

Abb. 3 Unterschiedliche Frageformen. **A** Geschlossenes Antwortformat mit Einfachnennung, **B** Halboffenes Antwortformat mit Mehrfachnennung und **C** Ratingfrage mit vierstufiger Antwortskala

für welche sämtliche Ränge explizit ausformuliert sind: So könnte im Rahmen des Science Slams die Aussage „Der Inhalt des Science-Slam-Vortrags war unterhaltsam" über die vierstufige Skala von „trifft zu" über „trifft eher zu" und „trifft eher nicht zu" bis hin zu „trifft nicht zu" bewertet werden. Eine entsprechende Verbalisierung ist für Befragte insgesamt leichter verständlich als rein numerische Skalen (Hussy et al. 2013). Zu berücksichtigen ist hier jedoch, dass die Vergabe solcher „Labels" bei allen Antwortkategorien nur für Skalen mit bis zu sieben oder teilweise auch neun Abstufungen sinnvoll ist (Franzen 2019).

In diesem Zusammenhang kann auch die Überlegung angestellt werden, wie viele Antwortkategorien überhaupt zur Verfügung stehen sollen. Wenige Abstufungen können zu einer verringerten Reliabilität führen, während zu viele Abstufungen keine weiteren Vorteile bringen und dazu tendieren, befragte Personen zu überfordern (Franzen 2019). In der Praxis haben sich daher vier- bis neunstufige Ratingskalen bewährt (Hussy et al. 2013). Während eine gerade Anzahl an Antwortkategorien die Tendenz zu einem der Skalenenden erzwingt („trifft eher zu" oder „trifft eher nicht zu"), besitzen ungerade Anzahlen eine neutrale Option in der Mitte – diese muss allerdings nicht zwangsläufig sinnvoll zu interpretieren sein: Befragte können damit angeben wollen, dass ihre Einstellung ambivalent ist („sowohl als auch") oder sie keine konkrete Einstellung

hinsichtlich der Frage besitzen („ich weiß nicht“) (Hussy et al. 2013). Um diese inhaltliche Ungenauigkeit aufzulösen, kann eine separate Antwortkategorie „ich weiß nicht“ bereitgestellt werden – allerdings zeigt sich, dass das Vorhandensein dieser Kategorie häufig zu einer verringerten Auseinandersetzung mit der Frage führt, und Befragte trotz eigener Meinung lieber auf verwertbare Angaben verzichten (Franzen 2019).

Schlussendlich ist zu klären, ob die Beschriftungen einer Skala bipolar, also mit zwei inhaltlich entgegengesetzten Endpunkten (z. B. „sehr langweilig“ bis „sehr unterhaltsam“), oder unipolar, also mit einem einzigen – jedoch unterschiedlich gewerteten – Endpunkt (z. B. „trifft zu“ bis „trifft nicht zu“), gewählt wird. Unipolare Skalen wie diese haben den Vorteil, dass ihre Dimensionen von Befragten nicht für jedes Item neu erfasst werden müssen, was die Befragung einfacher macht. In der Regel kommen Befragte damit besser zurecht (Franzen 2019).

4 Allgemeine Regeln zur Fragenkonstruktion

Bevor eine befragte Person verlässlich und wahrheitsgemäß antworten kann, muss von ihr zunächst sowohl die Semantik des Fragentextes als auch die Intention der Forschenden richtig interpretiert werden (Strack und Martin 1987). Wie gut Befragten dieser Prozess gelingt, bestimmt nicht zuletzt die Art und Weise, wie die gestellten Fragen formuliert wurden (Porst 2014). Die Qualität einer Frage hängt also unter anderem von der gewählten Formulierung ab – doch welche Aspekte sind hierfür zu berücksichtigen? Obwohl sich für die Fragenkonstruktion keine endgültigen Regeln aufstellen lassen, hat sich doch eine Art „Best Practice“ herauskristallisiert. Porst stellt in diesem Zusammenhang folgende zehn „Gebote“ auf (siehe Tab. 1).

Porsts Gebote beanspruchen allerdings keine Allgemeingültigkeit. Häufig kann es notwendig sein, gegen eines oder auch gegen mehrere Gebote zu verstoßen, wenn die konkrete Situation dies erfordert. So muss beispielsweise oft ein Kompromiss zwischen dem ersten, zweiten und zehnten Gebot gefunden werden, wenn Begriffe und Konzepte für das eindeutige Verständnis zunächst erklärt werden müssen. Die Gebote sind daher weniger als starre Regeln, sondern vielmehr als grundlegende „Wegweiser“ zu betrachten, welche für jede einzelne Frage erneut bedacht werden sollten (Porst 2014). So beschreibt Payne (1951) den Prozess der Fragenkonstruktion mehr als Kunst, denn als Wissenschaft – welcher erst durch individuelle Entscheidungen und Kreativität zu einer guten Frage führt.

Tab. 1 Zehn Gebote der Frageformulierung nach Porst (2014, S. 99 f.) – mit Erläuterungen

1. Du sollst einfache, unzweideutige Begriffe verwenden, die von allen Befragten in gleicher Weise verstanden werden!	… sonst sind Ergebnisse u. U. nicht vergleichbar
2. Du sollst lange und komplexe Fragen vermeiden!	… sonst werden Fragen schnell unverständlich und dafür anfällig, eines der restlichen Gebote zu brechen
3. Du sollst hypothetische Fragen vermeiden!	… sonst können Fragen nicht verlässlich beantwortet werden, da sie u. U. zu weit von der Lebensrealität Befragter entfernt sind
4. Du sollst doppelte Stimuli [d. h. zwei Befragungsgegenstände in einer Frage] und Verneinungen vermeiden!	… sonst können Fragen nicht wahrheitsgemäß beantwortet werden, da die Stimuli u. U. unterschiedlich beantwortet werden müssten oder Fragen werden bei insbesondere doppelter Verneinung schlicht missverstanden
5. Du sollst Unterstellungen und suggestive Fragen vermeiden!	… sonst können Fragen nicht wahrheitsgemäß beantwortet werden, da die Unterstellung u. U. nicht zutrifft oder Fragen können nicht verlässlich beantwortet werden, da u. U. eine Beeinflussung durch die Suggestion stattfindet
6. Du sollst Fragen vermeiden, die auf Informationen abzielen, über die viele Befragte mutmaßlich nicht verfügen!	… sonst können Fragen u. U. nicht wahrheitsgemäß beantwortet werden
7. Du sollst Fragen mit eindeutigem zeitlichen Bezug verwenden!	… sonst können Fragen nicht verlässlich beantwortet werden, da vage Zeitbezüge u. U. unterschiedlich verstanden werden
8. Du sollst Antwortkategorien verwenden, die erschöpfend und disjunkt (überschneidungsfrei) sind!	… sonst können Fragen nicht wahrheitsgemäß oder eindeutig beantwortet werden
9. Du sollst sicherstellen, dass der Kontext einer Frage sich nicht (unkontrolliert) auf deren Beantwortung auswirkt!	… sonst können Fragen nicht verlässlich beantwortet werden, da die vorangehende Frage oder Instruktion u. U. Einfluss auf die Beantwortung besitzt
10. Du sollst unklare Begriffe definieren!	… sonst können Fragen nicht verlässlich oder wahrheitsgemäß beantwortet werden

Abschließend sei darauf hingewiesen, dass die landläufige Meinung, Befragungen seien einfach zu konzipieren und umzusetzen, in Anbetracht der obigen Ausführungen mit Vorsicht zu genießen ist. Tatsächlich gibt es eine große Diskrepanz zwischen Fragen, die man im Alltag stellt, und soliden wissenschaftlichen Befragungen, wie sie im Rahmen wissenschaftskommunikativer Evaluationen zum Einsatz kommen sollten. Wissenschaftliche Befragungen erfordern somit eine gewissenhafte Planung und Vorbereitung – richtig umgesetzt ermöglichen sie im Rahmen der quantitativen Forschung allerdings auch die effiziente Erhebung beziehungsweise Auswertung großer Mengen an Informationen und schaffen die Vergleichbarkeit von Ergebnissen.

Literatur

Böhmert C, Niemann P, Hansen-Schirra S, Nitzke J (2021) Wen verstehen wir besser? Eine vergleichende Rezeptionsstudie zu Kurzmeldungen von Journalisten und Wissenschaftlern. In: Milde J, Welzenbach-Vogel IC, Dern M (Hrsg) Intention und Rezeption von Wissenschaftskommunikation. Halem, Köln, S 110–126

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

Döring N, Bortz J (Hrsg) (2016) Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften. Springer, Berlin (Springer-Lehrbuch)

Faulbaum F, Prüfer O, Rexroth M (2009) Was ist eine gute Frage? VS Verlag, Wiesbaden

Franzen A (2019) Antwortskalen in standardisierten Befragungen. In: Baur N, Blasius J (Hrsg) Handbuch Methoden der empirischen Sozialforschung. Springer Fachmedien, Wiesbaden, S 843–854

Fuchs M (2019) Mode-Effekte. In: Baur N, Blasius J (Hrsg) Handbuch Methoden der empirischen Sozialforschung. Springer Fachmedien, Wiesbaden, S 735–744

Gäde JC, Karin S-E, Werner C (2020) Klassische Methoden der Reliabilitätsschätzung. In: Moosbrugger H, Kelava A (Hrsg) Testtheorie und Fragebogenkonstruktion. Springer, Berlin Heidelberg, S 305–334

Hussy W, Schreier M, Echterhoff G (Hrsg) (2013) Forschungsmethoden in Psychologie und Sozialwissenschaften für Bachelor. Springer, Berlin (Springer-Lehrbuch)

Langer I, Schulz von Thun F, Tausch R (2019) Sich verständlich ausdrücken, 11. Aufl. Ernst Reinhardt Verlag, München.

Moosbrugger H, Kelava A (2020) Qualitätsanforderungen an Tests und Fragebogen („Gütekriterien"). In: Moosbrugger H, Kelava A (Hrsg) Testtheorie und Fragebogenkonstruktion. Springer, Berlin, S 13–38

Payne SL (1951) The Art of Asking Questions. University Press, Princeton

Porst R (2014) Fragebogen. Springer Fachmedien, Wiesbaden

Raithel J (2008) Quantitative Forschung. Ein Praxiskurs, 2., durchgesehene Aufl. VS Verlag (Lehrbuch), Wiesbaden.

Rasch B, Friese M, Hofmann W, Naumann E (2009) Quantitative Methoden 1. Einführung in die Statistik für Psychologen und Sozialwissenschaftler, 3. Aufl. Springer, Heidelberg
Reinecke J (2019) Grundlagen der standardisierten Befragung. In: Baur N, Blasius J (Hrsg) Handbuch Methoden der empirischen Sozialforschung. Springer Fachmedien, Wiesbaden, S 717–734
Strack F, Martin LL (1987) Thinking, judging, and communicating: A process account of context effects in attitude surveys. In: Hippler HJ, Schwarz N, Sudman S (Hrsg) Social information processing and survey methodology. Springer, New York (Recent Research in Psychology), S 123–148
Wirtz MA, Nachtigall C (2012) Deskriptive Statistik. Statistische Methoden für Psychologen Teil 1. 6., überarb. Aufl. Beltz Juventa, Weinheim

Christoph Böhmert ist Professor für Kommunikationspsychologie an der IU Internationale Hochschule im Fernstudium. Im Feld der Wissenschaftskommunikation erforscht er schwerpunktmäßig die Kommunikation von Risiken so wie von Zahlen und Statistiken.

Ferdinand Abacioglu, M.Sc., ist Psychologe und derzeit als wissenschaftlicher Mitarbeiter an der IU Internationale Hochschule tätig. Dort übt er Lehrtätigkeiten im Studiengang Psychologie (B.Sc.) aus.