Musikpsychologie Bd. 22, 165–179, © Hogrefe Verlag, Göttingen 2012

Statistische Poweranalyse als Weg zu einer ‚kraftvolleren' Musikpsychologie im 21. Jahrhundert

Friedrich Platz, Reinhard Kopiez und Marco Lehmann

1 Das Experiment als grundlegende empirische Methode

Auf die Frage, welche primären Zielsetzungen der empirischen Forschung allgemein zugrunde liegen, können aus unserer Sicht vor allem drei Ziele genannt werden: Effekte zu identifizieren, zu quantifizieren und zu prognostizieren (vgl. Hager, 2004; Kopiez, 2010). Das erste Ziel wird unter anderem durch das Aufdecken von Wirkungen im psychologischen Experiment erreicht. Nach Bredenkamp (1969) ist das Experiment hierbei durch mindestens drei notwendige Bedingungen – „Herstellung, Manipulation (systematische Variation) und Kontrolle" (S. 332) – definiert.

In diesem Zusammenhang wird unter *Herstellung* das Herbeiführen bzw. Realisieren theorierelevanter Experimentalbedingungen verstanden, unter denen ein zuvor durch Hypothesenbildung postulierter psychologischer Prozess beobachtbar sein soll. Dessen Stärke bzw. Vorhandensein wird dabei als Merkmalsausprägung einer Zielvariablen (abhängige Variable, kurz AV) verstanden.

Die Herstellungsphase folgt dabei einer Induktionslogik der Selektion: Hierunter verstehen wir das Ausschließen Theorie-irrelevanter Merkmale oder Zustände. Somit kann die Herstellung auch als konkrete Umsetzung eines durch die psychologische Theorie beschriebenen Systems definiert werden. Die häufig formulierte Kritik gegen das Experiment als empirische Methode ist, dass „die [hieraus gewonnenen] Daten nicht unbedingt für das *reale* Musikhören bzw. Musizieren zutreffend sein dürften" (Hemming, Busch & Auhagen, 2011, S. 36), da die Daten „künstlichen" Situationen entstammten. Diese Sichtweise resultiert aus dem Missverständnis, dass das Experiment eine Reduktion einer „realen" Hör- und Musiziersituation sei – wie auch immer in diesem Zusammenhang eine „reale" Situation von den Kritikern durch eine Theorie(-sprache) formalisiert wird. Diese Kritik ist insofern nicht zutreffend, als dass hierbei häufig die Alltagsrealität als Beurteilungskriterium für die Realitätsnähe einer experimentellen Situation dient. Dies kann allerdings kein Kriterium sein, denn naturwissenschaftlich-psychologische Theorien zeichnen nicht den Alltag in theoretischen Begriffen nach, sondern postulieren erst die Realität bestimmter psychologischer Prozesse und Entitäten. Die zum Beleg dieser Theorien entworfenen Experimente müssen dann natürlich nur dahingehend bewertet werden, wie sich in ihnen diese psychologischen Prozesse zeigen können.

Alle naturwissenschaftlichen Disziplinen arbeiten seit der Antike mit einer Wirklichkeit, die aus einer Theorie abgeleitet wird. Ein Beispiel hierfür ist die Atomtheorie: Kein Physiker hat bisher ein Atom gesehen, und dennoch operieren Wissenschaftler bis heute mit dieser theoretischen Einheit. Weiterhin setzen Atomphysiker sich nicht dem Zwang aus, Aussagen über das theoretische System hinaus leisten zu müssen, beispielsweise durch Annahmen darüber, welche Auswirkungen ein neues theoretisches Atomteilchen auf die Herzfrequenz des Menschen haben könnte, oder inwiefern der Teilchenbeschleuniger, in dem das Teilchen nachgewiesen wurde, jemals im Alltag der Nichtphysiker eine Rolle spielen könnte. Dementsprechend wird die psychologische Realität aus der psychologischen Theorie hergeleitet, wobei die Theorie als konstituierende Entität gesehen werden kann.

Die ausschließlich von der Experimentalleitung kontrollierte Veränderung mindestens eines latenten oder auch manifesten Faktors (unabhängige Variable, kurz UV) führt zu einer gewollten Variation *(Manipulation)* der experimentalen Ausgangsbedingung aus der Herstellungsphase. Neben der systematischen Variation der Bedingungen muss eine gleichzeitige *Kontrolle* der internen Validität erfolgen. Diese ist vor allem durch das Eliminieren von Störeinflüssen (Störvariablen) gekennzeichnet. Am wirkungsvollsten erfolgt eine Eliminierung durch die zufällige Zuweisung von Versuchspersonen zu Experimentalbedingungen. Man könnte auch formulieren, dass es sich bei der zufallsbestimmten Zuweisung um das Herzstück aller experimentellen Arbeit handelt. Nur hierdurch werden die Stufen der Störvariablen zufällig mit den verschiedenen Stufen der unabhängigen Variable kombiniert und damit kontrolliert (vgl. Huber, 2000, S. 67). Sollte diese Vorgehensweise nicht möglich sein (dies ist z. B. bei natürlichen Gruppen wie Schulklassen der Fall), muss dafür gesorgt werden, dass die Störvariable konstant gehalten wird. Dies ist bei quasi-experimentellen Bedingungen (wie bei Schulklassen) jedoch nicht immer möglich. In dieser Form geht sie als sogenannte Randbedingung ins Experiment ein. Ist eine Reduzierung der Störeinflüsse auf diesem Weg nicht möglich, müssen diese erhoben werden. Eine Kontrolle der Störvariablen erfolgt im Anschluss an das Experiment auf statistischem Weg durch Einbeziehen dieser Störvariablen als Kovariate in das statistische Gleichungssystem (vgl. Hager, 2004).

Aufgrund dieser Kontrolle der Störvariablen können sämtliche messbaren Unterschiede in der Zielvariablen (abhängige Variable, kurz AV) eindeutig kausal auf die systematische Manipulation (der UV) der Erhebungssituation zurückgeführt werden. Diese Experimentallogik ist auch als *ceteris paribus*-Regel bekannt (übers. „wobei alles andere gleich bleibt"). Sie schafft die Voraussetzung für die eindeutige Zuordnung des empirischen Unterschieds *(D)* zur experimentellen Manipulation. Dieser Mittelwertsunterschied wird als bedingte Auftrittswahrscheinlichkeit $p(D \mid H0)$[1] in Form eines inferenzstatistischen Tests bestimmt. Hierdurch wird das dritte Ziel empirischer Forschung erfüllt, die Vorhersage von Effekten. Ist eine Kausalität empirisch belegt worden, können

1 $p(D \mid H0)$ bezeichnet die Auftrittswahrscheinlichkeit *(p)* für den empirischen Mittelwertsunterschied *(D)* unter der Annahme, dass die Nullhypothese (H0) gilt (vgl. Sedlmeier & Renkewitz, 2008, S. 309 ff.).

unter Kenntnis der Bedingungsvoraussetzungen psychologische Prozesse vorhergesagt werden.

Die Frage, ob ein Wirkungsmechanismus vorliegt, wird mit der *statistischen Signifikanz* beantwortet. Fällt diese positiv aus, ist ein Effekt lediglich identifiziert, nicht jedoch quantifiziert, auch wenn Formulierungen wie „hoch signifikant" oder „höchst signifikant" Bedeutsamkeit nahelegen. Sie stellen den persuasiven Versuch dar, eine Effektgröße zu suggerieren, die mithilfe des Signifikanztests jedoch grundsätzlich nicht darstellbar ist. Nach Sedlmeier (2009) ist das ausschließliche Inspizieren des *p*-Werts Teil des weitverbreiteten „Signifikanztestrituals", dessen Aussagekraft maßgeblich überschätzt wird. So bleibt durch die fehlende Effektgrößenbestimmung die Frage nach der *Relevanz* eines Effekts häufig unbeantwortet.

Diese weitverbreitete Gewohnheit, lediglich Effekte ohne deren standardisierte Größenordnung zu publizieren, ist der Anlass dafür, in diesem Beitrag über die statistische a priori-Testplanung und deren Vorteile einen einführenden Überblick zu geben. Dabei stehen die Effekt- und Teststärkeberechnung, auch Poweranalyse genannt, im Mittelpunkt. Unser Hauptinteresse liegt weniger auf einer mathematisch orientierten Herleitung als auf einer Verständnisvermittlung der grundlegenden Vorteile dieser Herangehensweise. Pointiert formuliert sind wir der Überzeugung, dass durch diese Routinen der empirischen Musikforschung für Forscher die Wahrscheinlichkeit der Annahme von Forschungsbeiträgen steigt, die im internationalen Peer-Review-Prozess erfolgreich sind. Im Folgenden werden die wichtigsten Schritte, die aus dieser Forschungslogik hervorgehen, dargestellt. Eine zentrale Rolle wird hierbei die Effektgröße spielen.

2 Die Effektgröße *d* nach Cohen (1988)

2.1 Forderung 1: Abstände statt Unterschiede

Die Popularität des Signifikanztestrituals (Sedlmeier, 2009) geht auf das statistische Testen nach Fisher (1925) zurück. Cohen sieht in der einfach wirkenden Handhabung von Signifikanztests den Hauptgrund für die aus seiner Sicht wenig überraschende weite Verbreitung:

> „[Fisher's ideas] offered a deterministic scheme, mechanical and objective, independent of content, and led to clear-cut yes-no decisions" (Cohen, 1990, S. 1307).

Die Lösung des Problems, dass mithilfe des statistischen Testens nach Fisher (1925) lediglich Unterschiede durch das Bestimmen von Auftrittswahrscheinlichkeiten aufgedeckt werden konnten, besteht für Cohen in der Entwicklung eines Maßes, das vor allem *Abstände* oder *Zusammenhänge*[2] darstellen kann.

2 Im folgenden Text konzentrieren wir uns ausschließlich auf die Darstellung von Effektgrößen als Abstandsmaße. Gute Übersichten über vielfältige Effektgrößenmaße finden sich bei Kirk (2003), Ellis (2010) und ausführlich bei Cooper, Hedges & Valentine (2009).

Als Alternative zu der diffus wirkenden Aussage, dass eine Unterschiedlichkeit zwischen den empirischen Mittelwerten vorliegt (oder nicht vorliegt), kann Cohen durch das Angeben von Abständen die Größenordnung der Unterschiedlichkeit quantifizieren. Diese Überlegungen führen zur vorläufigen unstandardisierten Effektgrößengleichung (1), in der zunächst nur die Differenz *(D)* zweier (Roh-)Mittelwerte (die Mittelwerte von Experimental- und Kontrollgruppe) dargestellt ist.

$$D = \bar{x}_E - \bar{x}_K \qquad \text{Gleichung (1)}$$

2.2 Forderung 2: Bedeutungsvolle Skalierung der Abstände

Das Verwenden einer unstandardisierten Mittelwertsdifferenz *D* ist nur sinnvoll, wenn ihr Index bzw. ihre Bezugsgröße eine bedeutungsvolle Maßeinheit ist (z.B. km/h). Der gegenteilige Fall ist in Abbildung 1 dargestellt, in der die Mittelwertsdifferenz *D* in einer arbiträren Maßeinheit erfolgt.

Die Vergleichbarkeit von Effektgrößen unterschiedlicher Messinstrumente bzw. Studien ist erst mit einer einhergehenden Standardisierung der Mittelwertsdifferenz *D* an der gemeinsamen Merkmalsstreuung beider Gruppen (s_{pooled}) möglich. So formuliert Cohen die skalenunabhängige, vergleichbare Effektgröße *d* nach Gleichung (2) als Abstand zweier unabhängiger, an Standardabweichungen normierter Gruppenmittelwerte.

$$d = \frac{D}{s_{pooled}} = \frac{\bar{x}_E - \bar{x}_K}{\sqrt{\frac{(n_1 - 1)\cdot s_E^2 + (n_2 - 1)\cdot s_K^2}{n_1 + n_2 - 2}}} \qquad \text{Gleichung (2)}$$

Mit der Einführung einer Standardisierung von Abstandsmaßen geht die zusätzliche Möglichkeit einher, mithilfe von Konvertierungsformeln verschiedene Effektgrößen ineinander zu überführen (vgl. Cooper, Hedges & Valentine, 2009, S. 231 ff.).

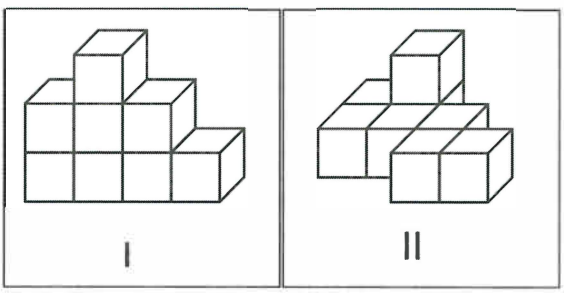

Abb. 1:
Versuch, Leistungsunterschiede in Rotationsaufgaben durch Mittelwertsdifferenz arbiträrer Einheiten (Anzahl „Rotationskörper") zu vergleichen (Beispielitem aus Birkel, Schein & Schumann, 2002).

2.3 Forderung 3: Aufdecken von Mindesteffekten – Relevanz statt Signifikanz

Um die Relevanz gefundener Effektgrößen besser bewerten zu können, gibt Cohen Konventionen (sogenannte Benchmarks) für die Bedeutsamkeit von Effektgrößen in der sozialpsychologischen Forschung an (1988, S. 24ff.). So unterscheidet Cohen zwischen einem kleinen ($d=0{,}20$), mittleren ($d=0{,}50$) und großen Effekt ($d=0{,}80$) für Mittelwertsunterschiede bei unabhängigen Gruppen. Metaphorisch beschreibt Cohen einen mittleren Effekt als einen Unterschied, der ‚mit bloßem Auge bereits zu beobachten ist' (vgl. Cohen, 1988, S. 26). Ein kleiner Effekt ist hingegen dann zu erwarten, wenn Wirkungen nicht unter Laborbedingungen gemessen werden und dadurch Störeinflüsse die maximale Wirksamkeit reduzieren bzw. der Messfehler größer als die wahre Merkmalsausprägung ist.

Die von Cohen genannten Bezugspunkte sollten jedoch nur als ungefähre Richtwerte gesehen werden (vgl. Rosnow & Rosenthal, 1989; Prentice & Miller, 1992). Je nach Forschungsdisziplin ist $d=0{,}20$ nicht immer mit einem inhaltlich „kleinen" Effekt gleichzusetzen. Eine alternative Betrachtung ist eine der prozentualen gemeinsamen Fläche beider gruppeneigener Merkmalsverteilungen. Je größer d ist, desto kleiner wird die gemeinsame Fläche, da die standardisierten Mittelwertsdistanzen größer werden und die Verteilungen sich immer weniger überlappen (vgl. Tabelle 1, zur Berechnung der gemeinsamen relativen Überschneidungsfläche siehe Bortz & Döring, 2006, S. 608). Ein historisches Beispiel für diese Art der Effektdarstellung zur Wirksamkeit psychotherapeutischer Behandlung findet sich bei Smith und Glass (1977).

Tab. 1:
Überlappungsbereich der Merkmalsverteilungen zweier Gruppen bei verschieden großen Effektgrößen

Effektgrößen	d	Überlappung (%)
Kleiner Effekt	0,20	92
Mittlerer Effekt	0,50	80
Großer Effekt	0,80	68

Anmerkung: Angenommen wird eine normalverteilte und varianzgleiche Merkmalsausprägung in beiden Gruppen (nach Bortz & Döring, 2006, S. 608).

3 Effektgrößenberechnung aus publizierten Daten

Für die eigene experimentelle Planung ist es notwendig zu wissen, wie groß die zu erwartende Effektgröße sein wird. Eine Schätzung sollte auf Grundlage publizierter Effektgrößenangaben erfolgen. Obwohl die APA-Standards (American Psychological Association, 2001, S. 25) vorschreiben, neben p-Werten auch

Effektgrößen der gefundenen Ergebnisse anzugeben, wird diese Forderung bis heute weitgehend ignoriert und von Reviewern auch nicht offensiv eingefordert. Dennoch können in vielen Fällen die deskriptiven Statistiken herangezogen werden, um Effektgrößen a posteriori zu berechnen. Diese können eine große Hilfe bei der Bewertung von Forschungsergebnissen sein. Selbst bei unvollständigen Angaben lassen sich Effekte noch weitestgehend ermitteln (vgl. Seifert, 1991; Cooper, Hedges & Valentine, 2009).

Als Beispiel soll die Effektgröße aus einer Klassifikationsstudie von Lamont und Webb (2010) berechnet werden, die mittels einer Tagebuchstudie zwei Musiknutzer- bzw. Hörertypen identifiziert haben wollen. So zeichnen sich die „Magpies" (die „Elstern") dadurch aus, dass sie selektiv bzw. nur wenige Musikstücke hören, während hingegen die „Squirrels" (die „Eichhörnchen") als Stückesammler bzw. Vielhörer beschrieben werden. Die Hypothese lautet, dass „Squirrels" in einem Hörtagebuch im Durchschnitt mehr Lieblingsstücke pro Woche nennen als die „Magpies". Nach Lamont und Webb (2010) liegt ein statistisch signifikanter Unterschied vor (Mann-Whitney $U = 1{,}5$, $p = 0{,}016$ (einseitig), $df = 7$).

Tab. 2:
Durchschnittliche Nennung gehörter Stücke pro Woche bei zwei postulierten Hörertypen

Hörtyp	*M*	*SD*	n
„Magpie"	8,0	4,69	4
„Squirrel"	13,8	1,92	5

Anmerkung: Daten entnommen aus Lamont & Webb (2010, S. 229 f.)

Um die Größe des Effekts zu berechnen, setzen wir die Werte aus Tabelle 2 in Gleichung 2 ein:

$$d = \frac{D}{s_{pooled}} = \frac{\bar{x}_E - \bar{x}_K}{\sqrt{\frac{(n_1 - 1)\cdot s_E^2 + (n_2 - 1)\cdot s_K^2}{n_1 + n_2 - 2}}} = \frac{13{,}8 - 8{,}0}{\sqrt{\frac{(5-1)\cdot 1{,}92^2 + (4-1)\cdot 4{,}69^2}{5+4-2}}} = 1{,}71$$

Der Effekt scheint nicht nur statistisch signifikant, sondern auch von praktischer Bedeutung zu sein (vgl. Ellis, 2010, S. 3). Sollte hier wohlmöglich ein ‚Mega-Effekt' gefunden worden sein? Ein Effekt, der mehr als nur ‚mit bloßem Auge' erkennbar ist? Bei näherer Betrachtung der geringen Versuchspersonenanzahl ($N = 9$) scheint das Ergebnis doch eher zweifelhaft zu sein. Die Lösung dieses Problems ist, dass hier die Effektgröße als Punktschätzer für einen Populationseffekt dient. Entsprechend der Forderung im *APA Publication Manual* (American Psychological Association, 2001, S. 22), sollte ein zusätzlicher Intervallschätzer (Standardfehler, 95 % Konfidenzintervall o. Ä.) genannt werden, um die Präzision des Punktschätzers einordnen zu können. Um der Forderung nachzu-

kommen, neben einem Punkt- auch einen Intervallschätzer (95 % Konfidenzintervall) anzugeben, muss zuerst der Standardfehler der Effektgröße für unabhängige Stichproben nach Gleichung (3) berechnet werden (vgl. Cooper, Hedges & Valentine, 2009).

$$SE_d = \sqrt{v_d} = \sqrt{\frac{n_E + n_K}{n_E \cdot n_K} + \frac{d^2}{2 \cdot (n_E + n_K)}} \qquad \text{Gleichung (3)}$$

Anschließend kann das 95 %-Konfidenzintervall (CI) nach Gleichung (4) berechnet werden[3].

$$CI_{95} = d \pm SE_d \cdot 1{,}96 \qquad \text{Gleichung (4)}$$

$$CI_{95} = 1{,}71 \pm 0{,}78 \cdot 1{,}96 = 1{,}71 \pm 1{,}528$$

Das Ergebnis ist wenig verblüffend: Aufgrund der geringen Stichprobengröße ist der Standardfehler der Effektgröße *(SE_d)* äußerst groß, weshalb das 95 %-Konfidenzintervall sich über mehrere Standardabweichungen erstreckt ($CI_u = 0{,}18$; $CI_o = 3{,}24$). Die unpräzise Effektschätzung ergibt, dass möglicherweise ein ‚sehr kleiner' oder aber ein ‚sehr großer' Effekt vorliegt ($d = 1{,}71$ [0,18; 3,24]). Ein derartiges – obwohl statistisch signifikantes – Ergebnis erscheint in Hinblick auf die unpräzise Schätzung kaum sinnvoll interpretierbar.

*3.1 Übung: Effektgrößenberechnung mit G*Power*

Mithilfe der freien Software G*Power[4] (Faul, Erdfelder, Buchner & Lang, 2009) ist es möglich, für viele unterschiedliche Auswertungsdesigns (a priori oder post hoc) Effektgrößen berechnen zu lassen. Um die im Text vorgestellte post hoc-Effektgrößenberechnung der Daten aus Tabelle 2 mit G*Power zu überprüfen, muss nach dem Starten des Programms der richtige statistische Test (❶) ausgewählt werden (s. Abbildung 2). Danach muss entschieden werden, welche Analyse durchgeführt werden soll. Hier sollte als Einstellung *„Post hoc: Compute achieved power"* gewählt werden (❷). Der dritte Schritt besteht in der Eingabe der deskriptiven Werte aus Tabelle 2. Dafür wird die Eingabemaske für die Effektgrößenberechnung benötigt. Diese wird durch den *Determine*-Button (❸) bereitgestellt. Da eine unterschiedliche Stichprobengröße vorliegt, darf nur der obere Teil der Eingabemaske genutzt werden (❹). Hier müssen als erstes die Gruppenmittelwerte eingetragen werden (❺). Leider kann G*Power bei ungleichen Gruppengrößen nicht mit den Standardabweichungen der Gruppen rechnen. Hier muss s_{pooled} ‚von Hand' ausgerechnet und in das Feld SD *within each group*

3 An dieser Stelle möchten wir auf das R-Paket MBESS hinweisen, das in der Lage ist, Vertrauensintervalle für Kontraste und d aus ANOVA-Ergebnissen zu berechnen.

4 Das Programm kann kostenlos unter folgender Adresse bezogen werden: http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/

eingefügt werden (❻). Nach Eingabe der Werte erhält man die Effektgröße, indem man abschließend auf den Calculate-Button (❼) klickt. G*Power bestätigt die ‚von Hand' gerechnete Effektgröße $d = 1{,}71$ (❽).

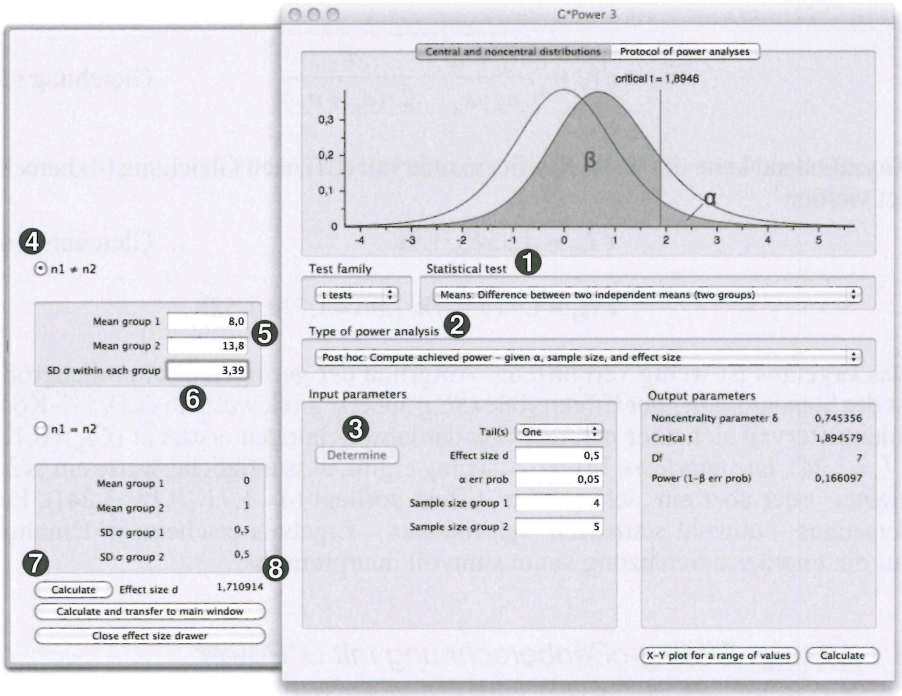

Abb. 2:
Screenshot der Benutzeroberfläche des Programms G*Power (Version 3) bei der Berechnung einer Effektgröße

4 Fehlertypen I und II

Ein weiteres zu berücksichtigendes Problem der empirischen Forschung sind die beiden grundsätzlichen Fehlertypen I und II. Mithilfe eines historischen Beispiels können die unterschiedlichen Fehlerarten bei der Interpretation eines statistischen Tests verdeutlicht werden (vgl. Ellis, 2010). Als die NASA 1979 in einer Pressemitteilung eine Aufnahme der Cydonia Region des Mars durch den NASA Viking 1 Orbiter als optische Täuschung darstellte (siehe Abbildung 3), wurde die Frage, ob es bereits menschliches Leben auf dem Mars gegeben hatte, wieder aufgenommen. In der sich neu entfachenden Diskussion bildeten sich ein Lager von Befürwortern und eins von Gegnern der Interpretation des Fotos als Abbild eines Kopfes. Die NASA verwies auf die Tatsache, dass es sich um eine optische Täuschung handele. Daher wurde festgestellt, dass dieses Bild kein Beweis für menschliches Leben auf dem Mars sei. Vertreter des Befürworter-Lagers wiederum vermuteten eine Verschwörung seitens der NASA und

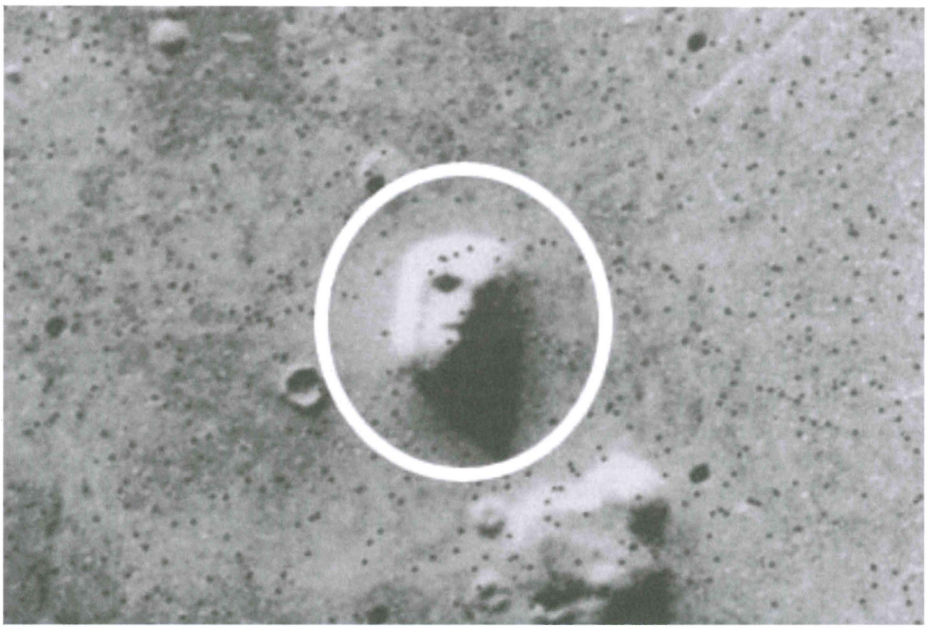

Abb. 3:
Aufnahme der Cydonia-Region des NASA Viking 1 Orbiter vom 25. Juli 1976. Diese optische Täuschung war für viele Menschen der Beweis, dass bereits menschenähnliche Lebewesen den Mars besucht hatten (aus Ellis, 2010, S. 51).

behaupteten, dass dieses Bild sehr wohl einen Beweis dafür darstelle, dass es menschliches Leben auf dem Mars gebe, zumindest gegeben habe.

Bei näherer Betrachtung der Aussagen beider Lager über das unverständliche Verhalten der jeweils anderen werden die beiden möglichen Fehlertypen deutlich: Aus Sicht der NASA nehmen die Befürworter einen Effekt an, der aus NASA-Sicht nicht vorhanden ist, während hingegen die Verschwörungstheoretiker der NASA vorwerfen, einen Effekt abzulehnen, obwohl dieser (augenscheinlich) vorhanden sei. Im ersten Fall handelt es sich um den Fehler der ersten Art (Typ I, falsch-positiv), *einen Effekt anzunehmen, obwohl dieser nicht vorhanden ist*. Demgegenüber handelt es sich bei dem Fehler der zweiten Art (Typ II, falsch-negativ) um die Entscheidung, *einen Effekt abzulehnen, obwohl dieser vorhanden ist.*

Beide Entscheidungsfehler können bei der Interpretation des statistischen Tests nie gleichzeitig ausgeschlossen, dafür aber ‚kontrolliert' werden, wenn diese als *a priori-Konzepte* in die Experimentalplanung eingehen. Die Wahrscheinlichkeit, einen Typ I-Fehler zu begehen, wird durch Festlegen der Irrtumswahrscheinlichkeit α quantifiziert, während der Typ II-Fehler durch die statistische Größe β festgelegt wird. Häufig wird statt des Fehlertyps II *(β)* von Teststärke gesprochen. Unter der Teststärke (oder engl. *power*) wird dabei die Wahrscheinlichkeit verstanden, mit der ein Effekt mit a priori angenommener Größe aufge-

deckt wird, sofern er auch tatsächlich vorhanden ist. Diese Wahrscheinlichkeit stellt das Gegenereignis zum Fehlertyp II dar ($power = 1-\beta$).

Nach Fisher (1925) sollte die Wahrscheinlichkeit, einen Typ I-Fehler zu begehen, unter 5 % liegen. Dementsprechend bedeutet ein vorher festgelegtes Alpha-Niveau von $\alpha = 0{,}05$, dass bei 20 Schülern, die alle nach einem Effekt suchen, obwohl keiner vorhanden ist, lediglich ein Schüler sich lächerlich macht, weil er etwas sieht, was nicht da ist (vgl. Ellis, 2010, S. 49).

5 Die 5-20-Regel

Um der Gefahr eines Typ I-Fehlers aus dem Weg zu gehen, könnte der Alpha-Wert – ganz im Fisher'schen Sinn – äußerst klein definiert werden (z. B. $\alpha = 0{,}0001$). Mit dem Minimieren des kritischen Alpha-Werts sinkt jedoch auch gleichzeitig die Teststärke. Aufgrund unterschiedlicher Risikobewertungen für die Fehler vom Typ I und II muss diesem Dilemma der Fehlerminimierung bei der statistischen Planung Aufmerksamkeit geschenkt werden. Als „Daumenregel" ist die 5-20-Regel bekannt, nach der die Wahrscheinlichkeit eines Fehlers vom Typ I bei 5 % und vom Typ II bei 20 % liegen sollte (vgl. Ellis, 2010). Nach Cohen (1988, S. 5) ist bei dieser Konstellation das Risiko nur viermal höher, fälschlicherweise die Nullhypothese abzulehnen (Typ I) als fälschlicherweise den Effekt abzulehnen (Typ II).

Pointiert formuliert hat diese Vorgehensweise mit der Fokussierung auf den Alpha-Fehler (Typ I) dazu geführt, dass die meisten Zeitschriften durch die Tendenz zur Veröffentlichung falsch positiver Ergebnisse („etwas gefunden") gekennzeichnet sind. Solange der Typ II-Fehler nicht ausreichend berücksichtigt wird (z. B. durch eine genügend hohe Testpower), kann zusätzlich nicht abgeschätzt werden, wie häufig falsch negative Ergebnisse („nichts gefunden") veröffentlicht werden.

*5.1 Beispiel: Retrospektive Power-Analyse mithilfe von G*Power*

Warum ist nun eine Power-Analyse vor Beginn eines Experiments wichtig? In einem Gedankenexperiment soll angenommen werden, dass experimentell untersucht wurde, ob es Unterschiede in der Bewertung eines Musikbeispiels in Abhängigkeit von der Darbietungsform (audiovisuell vs. auditiv) gibt. Angenommen, insgesamt 50 Versuchspersonen wurden eingeladen, die zufällig einer der beiden Versuchsbedingungen zugewiesen wurden. Nehmen wir weiterhin an, dass nach Abschluss der Datenerhebungsphase der statistische Test ein nichtsignifikantes Ergebnis zeigte. Nach herkömmlichem Vorgehen (ohne a priori-Poweranalyse) würde nun angenommen werden, dass es keinen Bewertungsunterschied zwischen auditiv bzw. audiovisuell dargebotener Musik gegeben hat. Vor dem Hintergrund, dass keine a priori-Annahme über die Teststärke geleistet wurde, könnte sich mit der Entscheidung die Nullhypothese anzunehmen mit hoher Wahrscheinlichkeit ein Typ II-Fehler einstellen. Dieser Vorwurf soll mit-

hilfe einer post hoc-Poweranalyse mittels G*Power überprüft werden. Für die Parametereingabe wird angenommen, dass im Experiment ein mittlerer Effekt ($d = 0{,}50$) aufgedeckt wurde.

Nach Eingabe der Werte in G*Power und dem Aktivieren der Option post hoc-Analyse *(Statistical Test)* erhält man als Ergebnis eine Testpower von 41 %. Anders formuliert, die Wahrscheinlichkeit einen Effekt abzulehnen, obwohl dieser vorhanden ist, beträgt 59 %! Vor diesem Hintergrund ist es empfehlenswert, sich für keine Ergebnisinterpretation zu entscheiden. Derartige Experimente werden häufig als ‚unterpowert' bezeichnet, da sie die Nullhypothese ‚bevorzugt' haben. Man spricht dabei auch häufig von der Beobachtung, dass der Effekt keine ‚faire' Chance hatte sich durchsetzen. Eine Studie gilt dagegen als ‚überpowert', wenn die Nullhypothese keine ‚faire' Chance hatte sich durchzusetzen. Aus diesem Dilemma resultiert zwangsläufig die Frage, wie man denn nun zur *richtigen* Stichprobengröße gelangt.

6 „Optimale" Stichprobengröße

Wie viele Versuchspersonen eingeladen werden müssen, ist für viele Forschende eine der gefürchtetsten Fragen. Nehmen wir die Frage, ob 30 Personen ausreichen oder ob doch lieber 300 eingeladen werden sollten. Unter der Annahme, dass die Versuchspersonen eine Aufwandsentschädigung von zehn Euro erhalten, würden sich die Kosten zwischen 300 und 3 000 Euro bewegen. Damit wird die richtige Stichprobenberechnung zu einer Frage des verantwortlichen Umgangs mit Ressourcen. Die Frage nach der optimalen Stichprobe ist an die Frage der Testpower, der Toleranz für den Fehlertyp I und der angenommen Effektgröße geknüpft. Erst mit diesen drei Informationen kann die Gesamtstichprobengröße ermittelt werden. Liegen keine Ergebnisse aus früheren Studien oder aus Meta-Analysen vor (s. beispielsweise Platz & Kopiez, 2012), können mit G*Power alternative Szenarien erstellt werden.

*6.1 Stichprobengrößenszenarien als Basis für die Stichprobenplanung mit G*Power*

Nach dem Starten des Programms wird der Szenario-Modus durch Aktivieren der Option *X-Y plot for a range of values-Button* gestartet. Ein zweites Fenster erscheint, in dessen unterem Teil die Dateneingabe erfolgt. Da die Stichprobengröße von Interesse ist, sollte diese auf der Y-Achse als Funktion einer sich verändernden Effektgröße abgetragen werden. In diesem Szenario wird sowohl das Alpha- als auch das Beta-Niveau entsprechend der 5-20-Regel konstant gehalten. Das Ergebnis ist in Abbildung 4 dargestellt. Ihr kann entnommen werden, dass bei einem Alpha-Niveau von 5 % und einer Wahrscheinlichkeit von 80 % erst mit insgesamt 788 Versuchspersonen eine kleine Effektgröße ($d = 0{,}20$) aufgedeckt werden kann, sofern diese auch tatsächlich vorhanden ist. Je größer die zu erwartende Effektgröße ist, desto geringer wird die Stichproben-

größe, mit der die Effektgröße bei einer Wahrscheinlichkeit von 80 % zu einem signifikanten Ergebnis führt. Unter dem Karteireiter *Table* findet man eine tabellarische Auflistung der genauen Kennwerte.

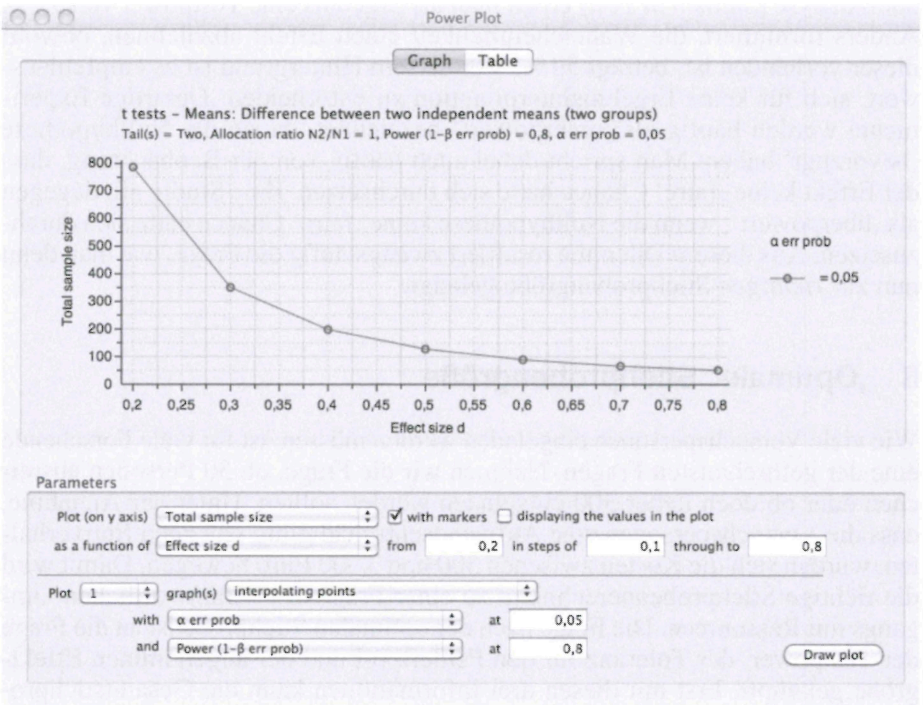

Abb. 4:
Die Oberfläche der Software G*Power visualisiert unterschiedliche Stichprobengrößen als Funktion verschiedener Effektgrößen unter Konstanthalten von α und power (1–β) in Form eines Liniendiagramms.

Bleibt als Vorgehensweise bei der Experimentalplanung nur das Schätzen der Stichprobengröße übrig, weil Werte aus vorherigen Studienergebnissen fehlen, ist es empfehlenswert, eher konservativ vorzugehen und nicht eine Punkt-, sondern eine Intervallschätzung vorzunehmen. Dabei sollte als Leitfrage immer formuliert werden, von welchem Mindesteffekt man im ungünstigsten Fall ausgehen möchte. In der Experimentalplanung muss also festgelegt werden, welche Mindesteffektgröße als forschungsrelevant angesehen wird. Eine einfache Daumenregel wie „viel hilft viel" kann dazu führen, dass entweder geringe und vielleicht unbedeutende Effekt aufgedeckt werden oder dass Ressourcen verschwendet werden, weil eine weitere Vergrößerung der Stichprobe kaum Zuwachs an Testpower bringt. Besonders dann, wenn Hilfskräfte für die Dateneingabe benötigt werden, ist demnach die Frage nach der optimalen Größe und nicht die nach der maximal möglichen Größe einer Stichprobe entscheidend.

Je kleiner der a priori angenommene Effekt ist, desto größer also ist die benötigte Gesamtstichprobe. Gleichermaßen wichtig ist die genaue und vielschichtige Kontrolle der internen Validität eines Experiments vor allem bei kleinen Effektgrößen. Erst sie ermöglicht die größtmögliche Entfaltung der Effektgröße.

7 Zusammenfassung und Ausblick

Das a priori-Konzept der Poweranalyse und Stichprobenplanung kann einen Beitrag zur Vermeidung „aussichtsloser" Forschung leisten, da vor allem kleine, durch gute Stichprobenumfangsplanung abgesicherte Effekte experimentell untersucht werden können. Durch die Effektgrößenbestimmung ist es darüber hinaus möglich, experimentell begründete Unterschiede über Studien hinweg vergleichbar zu machen. Auf der Basis von Effektgrößenangaben mit zusätzlichen Konfidenzintervallen und genügend hoher Testpower werden Effekte der eigenen Forschung mit hoher Wahrscheinlichkeit auch in Replikationsstudien auftreten.

Ein weiterer Vorteil bei der Angabe von Effektgrößen und deren Konfidenzintervallen ist das gleichzeitige Vermeiden von „strong claims" (s. hierzu ausführlich Swales & Feak, 2004, S. 112). Solche Behauptungen werden bis heute in der Regel ausschließlich auf Basis eines *p*-Werts publiziert. Es sei noch einmal darauf hingewiesen, dass der *p*-Wert deshalb problematisch ist, weil sowohl die Stichprobengröße als auch die „tatsächliche" Effektgröße partielle Einflussgrößen des *p*-Werts sind, was nichts anderes als eine Konfundierung ist. Somit kann der *p*-Wert als einzige Angabe zur Effektgröße nicht verwendet werden.

Studien, die ihre Ergebnisse in Effektgrößen angeben, können darüber hinaus in einer Meta-Analyse zusammengefasst werden. Mit ihr ist es möglich, im Rahmen von Theoriebildung weitere Voraussetzungen für die Sichtbarkeit des Effekts durch Meta-Regressionen zu bestimmen. Darüber hinaus erhöht sich durch Aggregieren vieler Studien in der Regel die Power und die Genauigkeit, mit der ein Effekt geschätzt werden kann (vgl. Cohn & Becker, 2003).

7.1 Die Leitidee der Daten-Nachhaltigkeit

Damit die eigenen Forschungsergebnisse für andere nützlich sind, sollte die Datenaufbereitung nicht vernachlässigt werden. Nach Cortina und Hossein (2000) sind beim Publizieren empirischer Studien vor allem vier Punkte unbedingt zu beachten:

1. Vollständige Dokumentation sämtlicher deskriptiven Angaben für sämtliche Bedingungen: vor allem Mittelwert, Standardabweichungen und Zellbesetzung.
2. Angabe der Korrelationen zwischen den Messzeitpunkten bei messwiederholten Designs, denn die Effektgrößen können ansonsten nicht präzise geschätzt werden.

3. Angaben zur a priori angenommen Testpower zwecks Vermeidung von *Unterpowerung*.
4. Angabe von Effektgrößen mit Konfidenzintervallen für alle beobachteten Effekte.

7.2 Leseempfehlung und Software zur Bestimmung von Effektgrößen und zur Durchführung von Meta-Analysen

Abschließend empfehlen wir neben Sedlmeier und Renkewitz (2008) die hervorragende englischsprachige Einführung von Ellis (2010). Dieses Buch ist eine tiefergehende Einführung in die Gesamtthematik, wobei auf eine verständliche Darstellung viel Wert gelegt wurde. Darüber hinaus empfehlen wir das detailreiche Handbuch von Cooper, Hedges und Valentine (2009), in dem weitere Effektgrößenbestimmungstechniken für unvollständig publizierte Daten aufgeführt sind. Die Autoren legen dabei den Schwerpunkt dieses Handbuchs in die Erstellung einer Meta-Analyse und führen den Leser Schritt für Schritt durch deren Phasen, angefangen von der Literatursuche bis hin zu Formen der Datenvisualisierung. Darüber hinaus werden mathematische Zusammenhänge dargestellt.

Als Softwareunterstützung empfehlen wir als internationalen Standard neben dem hier vorgestellten G*Power das kommerzielle Programm CMA (Comprehensive Meta-Analysis). Für die Statistikumgebung R wird das Paket *Metafor* empfohlen. Dieses hat den größten Umfang zur Berechnung von Effektgrößen. Darüber hinaus bietet *Metafor* umfangreiche Visualisierungsmöglichkeiten für aggregierte Daten an.

7.3 Schlussbetrachtung

Insgesamt sind wir der Überzeugung, dass wir mit der Einhaltung dieser Best practice-Regeln beim Forschen und Publizieren eine gute Chance haben, auch auf dem internationalen Niveau etablierter Disziplinen der Psychologie mithalten zu können.

Im Prinzip verhält es sich mit der empirischen Forschung wie mit der Musikaufnahme im Tonstudio: Die Signalqualität entsteht an der Quelle. Falsche Töne sind im Nachhinein nur bedingt und mit größten Mühen, in vielen Fällen jedoch überhaupt nicht mehr korrigierbar.

Literatur

American Psychological Association (2001). *Publication Manual* (5. Auflage). Washington: American Psychological Association.

Birkel, P., Schein, S. A. & Schumann, H. (2002). *Bausteine-Test (BST).* Göttingen: Hogrefe.

Bortz, J. & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler* (4. Auflage). Berlin: Springer.

Bredenkamp, J. (1969). Experiment und Feldexperiment. In C. Graumann (Hrsg.), *Handbuch der Psychologie* (1. Halbband: „Theorien und Methoden der Sozialpsychologie", Band 7, S. 332–374). Göttingen: Hogrefe.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2. Auflage). New Jersey, USA: Lawrence Erlbaum Associates, Inc.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45* (12), 1304–1312.

Cohn, L. & Becker, B. (2003). How meta-analysis increases statistical power. *Psychological Methods, 8* (3), 243–253.

Cooper, H., Hedges, L. & Valentine, J. (2009). *The handbook of research synthesis and meta-analysis.* New York: Russel Sage Foundation.

Cortina, J. & Hossein, N. (2000). *Effect size for ANOVA designs.* Thousand Oaks, CA: SAGE Publications, Inc.

Ellis, P. D. (2010). *The essential guide to effect sizes. Statistical power, meta-analysis, and the interpretation of research results.* Cambridge: Cambridge University Press.

Faul, F., Erdfelder, E., Buchner, A. & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41* (4), 1149–1160.

Fisher, R. (1925). *Statistical methods for research workers.* Edinburgh: Oliver & Boyd.

Hager, W. (2004). *Testplanung zur statistischen Prüfung psychologischer Hypothesen.* Göttingen: Hogrefe.

Hemming, J., Busch, V. & Auhagen, W. (2011). Grundlegende Verfahren der empirischen Sozialforschung. In W. Auhagen, V. Busch & J. Hemming (Hrsg.), *Systematische Musikwissenschaft* (S. 35–40). Laaber: Laaber.

Huber, O. (2000). *Das psychologische Experiment: Eine Einführung* (3. Auflage). Bern: Huber.

Kirk, R. E. (2003). The importance of effect magnitude. In S. Davis (Ed.), *Handbook of research methods in experimental psychology* (pp. 83–105). Oxford: Blackwell.

Kopiez, R. (2010). Empirische Methoden. In H. de la Motte-Haber, H. von Lösch, G. Rötter & C. Utz (Hrsg.), *Lexikon der systematischen Musikwissenschaft* (S. 95–99). Laaber: Laaber.

Lamont, A. & Webb, R. (2010). Short- and long-term musical preferences: What makes a favourite piece of music? *Psychology of Music, 38* (2), 225–241.

Platz, F. & Kopiez, R. (2012). When the eye listens: a meta-analysis of how audio-visual presentation enhances the appreciation of music performance. *Music Perception, 30* (1).

Prentice, D. & Miller, D. (1992). When small effects are impressive. *Psychological Bulletin, 112* (1), 160–164.

Rosnow, R. & Rosenthal, R. (1989). Procedures and the justification of knowledge in psychological science. *American Psychologist, 44,* 1276–1284.

Sedlmeier, P. (2009). Beyond the significance test ritual. *Zeitschrift für Psychologie, 217* (1), 1–5.

Sedlmeier, P. & Renkewitz, F. (2008). *Forschungsmethoden und Statistik in der Psychologie.* München: Pearson Studium.

Seifert, T. (1991). Determing effect sizes in various experimental designs. *Educational and Psychological Measurement, 51,* 341–347.

Smith, M. & Glass, G. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32* (9), 752–760.

Swales, J. M. & Feak, C. B. (2004). *Academic writing for graduate students: Essential tasks and skills* (2. ed.). Ann Arbor: University of Michigan Press.