SFU Forschungsbulletin SFU Research Bulletin 8/1 (2020), S. 87-105 DOI: 10.15135/2020.8.1.87-105

# Der t-, Welch- und U-Test im psychotherapiewissenschaftlichen Forschungskontext

## Empfehlungen für Anwendung und Interpretation

# T-, Welch- and U-test in psychotherapy science

## recommendations for application and interpretation

David Seistock, Anastasiya Bunina, Jan Aden

*Kurzzusammenfassung*

In diesem ersten Beitrag der Serie Statistik in der Psychotherapiewissenschaft wird die Anwendung des t-, Welch- sowie U-Tests bei unverbundenen Stichproben im Sinne eines Best-Practice Ansatzes vorgestellt. Neben Empfehlungen für eine (1) optimale Verfahrenswahl, (2) dem Einsatz von Effektstärken, (3) der Bestimmung der Ergebnisrelevanz sowie (4) der Vorstellung von Reportkonventionen für die Ergebnisdarstellung, wird vor allem (5) auf das Problemfeld der Zuverlässigkeit statistischer Entscheidungen im psychotherapiewissenschaftlichen Forschungskontext und (6) Möglichkeiten zur aktiven Einflussnahme durch die Forscher*innen, eingegangen.

*Schlüsselwörter*

t-Test, Welch-Test, U-Test, Reportkonventionen, Oversampling, Undersampling, Effektstärken, statistische Entscheidungen

© SFU Forschungsbulletin SFU Research Bulletin 2020 (1)

## *Abstract*

In this first contribution to the Statistics series in psychotherapy science, the application use of the t-, Welch- and U-test for unrelated samples is presented in the sense of a best practice approach.

In addition to recommendations for (1) the optimal choice of procedure, (2) the use of effect sizes, (3) the designation of relevant results and (4) report conventions for the presentation of results, (5) the problem of reliable statistical decision making in the research context of psychotherapy science is adressed and therefore, (6) suggestions for dealing with this potential problem are identified.



## *keywords*

t-Test, Welch-Test, U-Test, reporting conventions, Oversampling, Undersampling, statistical decision making

## *Einsatzfeld und Background*

*Lesehinweis: Falls Sie bestimmte inhaltliche Ausführungen (z.B. Formeln etc.) überspringen möchten, finden Sie die wichtigsten Punkte der jeweiligen Textpassage in den Anmerkungen am linken Seitenrand zusammengefasst. Außerdem finden Sie innerhalb des Textes Exkurse, in denen bestimmte im Fließtext erwähnte Sachverhalte näher erläutert werden.*

Verfahren zur Überprüfung statistisch relevanter Unterschiede stellen eine der am häufigsten angewandten Verfahrensgruppe sozial- und humanwissenschaftlicher Disziplinen dar, so auch in der quantitativen Psychotherapieforschung. Die möglichen Anwendungsfelder dieser Verfahrensklasse sind zahlreich: Lässt sich beispielsweise eine Verbesserung der Depressionswerte nach einer sechswöchigen Therapie feststellen, gehen mit dem Einsatz verschiedener Therapiemethoden bestimmte Veränderungen eines Störungsbildes einher, oder spielt das Geschlecht eine Rolle für die Ausprägung ausgewählter Merkmale (z.B. Stresslevel oder Schwere einer depressiven Symptomatik)? All jene Fragen können unter Zuhilfenahme statistischer Unterschiedstestungen untersucht werden. Die quantitative Forschungsmethodik stellt eine Vielzahl an Verfahren bereit, unter deren Anwendung Differenzen zwischen zwei oder mehr Gruppen/Stichproben beziehungsweise Messzeitpunkten überprüft werden können. Die Testung eines statistisch signifikanten Unterschieds zwischen zwei Gruppen/Stichproben/Messzeitpunkten und dessen Generalisierung stellt dabei eine in der Praxis häufig eingesetzte Auswertungsform dar.

*Bestimmung der Stichprobenform (verbunden vs. unverbunden)*

Unterschiedstestungen für zwei Gruppen[1] (Stichproben) lassen sich anhand der Form der zu untersuchenden Stichproben weitestgehend in zwei Methodenbereiche gliedern: Verfahren zur Bestimmung von signifikanten Unterschieden bei verbundenen sowie bei unverbundenen Stichproben. Ein klassisches Beispiel für verbundene Stichproben stellen Therapieevaluationen und Effektivitätsnachweise dar. Hierfür wird den Klient*innen in der Regel ein störungsspezifisches Messinstrument zu Beginn sowie nach Abschluss einer Therapie vorgegeben (z.B. das Beck-Depressions-Inventar vgl. Kühner, Bürger, Keller & Hautzinger, 2007) und überprüft, ob sich nach Abschluss der Therapie eine signifikante Veränderung in der Ausprägung der Depression beobachten lässt (siehe Tab. 1). Unverbundene Stichproben liegen überall dort vor, wo zwar ein Unterschied zwischen zwei Gruppen überprüft werden soll, die erhobenen bzw. gemessenen Werte jedoch nicht paarweise zugeordnet werden können. Beispielsweise könnte untersucht werden, ob sich ein Unterschied in der Qualität der Therapeut*innen-Klient*innen-Beziehung zwischen Therapieform A und Therapieform B zeigt. Hierfür könnte ein Fragebogen zur Evaluation der Qualität der Therapeuten-Klienten-Beziehung (z.B. Bonner Fragebogen für Therapie und Beratung (Berth & Brähler, 2003)) sowohl Personen, welche sich seit zwei Monaten in einer Therapie der Form A befinden, als auch welchen, die sich seit zwei Monaten in einer Therapie der Form B befinden, vorgegeben werden. Im Vergleich zum erläuterten Beispiel

[1] Die Begriffe „Gruppe“ und „Stichprobe“ werden im Rahmen dieses Beitrages synonym verwendet.

bei verbundenen Stichproben (Messung der BDI-Werte vor und nach der Therapie), werden hier die Qualitätseinschätzungen von zwei verschiedenen Personengruppen (Klient*innen der Therapieform A vs. Klient*innen der Therapieform B) zu nur einem Zeitpunkt erhoben (siehe Tab. 2). Somit lassen sich die gemessenen Werte nicht paarweise zuordnen, da pro Person nur ein Messwert vorliegt.

Tab.1: Bsp. für verbundene Stichproben

| | Messzeitpunkt 1 | Messzeitpunkt 2 |
|---|---|---|
| KlientIn A | 25 | 20 |
| KlientIn B | 75 | 60 |
| KlientIn C | 35 | 27 |
| … | … | … |

Tab.2: Bsp. für unverbundene Stichproben

| Therapieform A | | Therapieform B | |
|---|---|---|---|
| KlientIn A | 25 | KlientIn D | 60 |
| KlientIn B | 75 | KlientIn E | 75 |
| KlientIn C | 35 | KlientIn F | 80 |
| … | … | … | … |

*Verfahren für zwei unverbundene Stichproben: t-Test, Welch-Test, U-Test*

Um zu untersuchen, ob sich zwei unverbundene Gruppen statistisch signifikant in der Ausprägung eines mindestens rangskalierten (ordinalen) Merkmals unterscheiden, ist es üblich, in Abhängigkeit bestimmter Voraussetzungen, eines der folgenden Auswertungsverfahren heranzuziehen: den t-Test, den Welch-Test sowie den Mann-Whitney-U-Test. Da es sich bei den drei angeführten Tests um sogenannte inferenzstatistische Verfahren handelt, setzen diese die Formulierung eines Hypothesenpaares (Null- und Alternativhypothese) voraus, die für die Population verallgemeinert werden soll (schließende (inferenz) Statistik). Um das bereits erläuterte Beispiel zur Untersuchung von Unterschieden in der wahrgenommenen Qualität der Klient*innen-Therapeut*innen-Beziehung erneut aufzugreifen, wäre folgende Formulierung denkbar:

*Hypothesen-formulierung*

H0: Es besteht kein signifikanter Unterschied zwischen Klient*innen der Therapieform A und Klient*innen der Therapieform B hinsichtlich der wahrgenommenen Qualität der Klient*innen-Therapeut*innen-Beziehung

H1: Es besteht ein signifikanter Unterschied zwischen Klient*innen der Therapieform A und Klient*innen der Therapieform B hinsichtlich der wahrgenommenen Qualität der Klient*innen-Therapeut*innen-Beziehung.

Jedes der drei angeführten Verfahren (t-Test, Welch-Test, U-Test) verfügt über bestimmte Voraussetzungen (z.B. Skalenniveau des untersuchten Merkmals), welche erfüllt sein müssen, um eine korrekte Anwendung und Interpretation der erzielten Ergebnisse gewährleisten zu können. Die Verfahrenswahl hängt somit stark von der jeweiligen

Datenlage ab. Gemeinsam ist den angeführten Verfahren, die Verwendung von Prüfverteilungen. So wird im Rahmen einer bestimmten Prüfverteilung stets ein numerischer Verteilungswert berechnet, welcher die Größe des Unterschieds zwischen den beiden Gruppen in der Ausprägung des untersuchten Merkmals widerspiegelt. Diese Verfahren folgen somit einer einheitlichen Struktur, die sich anhand der Frage, wen (welche Gruppen z.B. Therapie A und B, Männer und Frauen) vergleiche ich hinsichtlich wessen[2] (untersuchtes Merkmal: z.B. Qualität der Klient*innen-Therapeut*innen-Beziehung, Extraversion, Angst), illustrieren lässt. Um zu bestimmen, ob es sich um einen statistisch bedeutsamen (systematischen/signifikanten) oder unbedeutenden (zufälligen) Unterschied handelt, wird der anhand der Prüfverteilung (abhängig vom jeweiligen Verfahren z.B. t-Verteilung beim t-Test) berechnete Verteilungswert in einen Wahrscheinlichkeitswert (p-Wert) überführt und mit einer festgelegten Signifikanzgrenze (in den Sozialwissenschaften in der Regel α= 5 %) verglichen. Ergibt sich ein Signifikanzwert (p-Wert) von 5% (p= .050) oder geringer (z.B. p = 2% /p= .02), handelt es sich um einen statistisch signifikanten Unterschied. Die Wahl der Signifikanzgrenze (Alpha-Niveau) hängt stark von der wissenschaftlichen Disziplin ab, im Rahmen derer die jeweilige Untersuchung stattfindet. So ist es beispielsweise bei manchen Fragestellungen in der humanmedizinischen Forschung durchaus üblich, ein geringeres Signifikanzniveau von beispielsweise 1% anzunehmen.

*Signifikanz-bestimmung*

***Exkurs: einseitige vs. zweiseitige Unterschiedstestungen***

*Je nach formulierter Hypothese können ein- oder zweiseitige Unterschiedstestungen herangezogen werden. Zweiseitige Unterschiedstestungen beziehen sich auf Hypothesen, die sich lediglich auf die Untersuchung eines Unterschiedes beziehen, ohne dabei eine konkrete Richtung anzunehmen (z.B. Unterscheiden sich SchichtarbeiterInnen von Nicht-SchichtarbeiterInnen hinsichtlich ihrer Beschwerden im Lebensbereich Familie? (z.B. mit BOSS erhoben, vgl. Hagemann & Geuenich, 2009)). Einseitige Testungen können hingegen nur dann angewandt werden, wenn bereits eine begründete Vorannahme bezüglich der Richtung des Unterschiedes besteht (z.B. Unterscheiden sich SchichtarbeiterInnen von Nicht-SchichtarbeiterInnen hinsichtlich ihrer Beschwerden im Lebensbereich Familie dahingehend, dass SchichtarbeiterInnen größere Beschwerden aufweisen?). Werden einseitige Testungen angewandt, muss die Richtung des Unterschiedes in der Formulierung der Hypothesen berücksichtigt werden (z.B. H0: Es besteht kein signifikanter Unterschied zwischen SchichtarbeiterInnen und Nicht-SchichtarbeiterInnen hinsichtlich ihrer Beschwerden im Lebensbereich Familie, dahingehend dass SchichtarbeiterInnen größere Beschwerden aufweisen).*

*Bei der Interpretation des Ergebnisses einer einseitigen Unterschiedstestung gilt es zwei zusätzliche Punkte zu beachten: Zum einen kann der Signifikanzwert (p-Wert) durch zwei dividiert werden, wenn einseitige Testungen nicht speziell in der Auswertungssoftware berücksichtigt werden können (z.B. Auswahloption „einseitige Testung"). Ergibt sich beispielsweise ein p-Wert von .070, kann dieser noch „halbiert" werden und würde mit p= .035 auf einen signifikanten Unterschied hindeuten.*

[2] Diese Formulierung entspricht einer alltagssprachlichen Diktion, die zur besseren Verständlichkeit verwendet wird.

*Zum anderen müssen die beiden Gruppen im Anschluss an ein signifikantes Ergebnis anhand deskriptiver Kennwerte (je nach Verfahren z.B. Mittelwert, Median) verglichen werden, um die, in der Hypothese spezifizierte Richtung des Unterschiedes (z.B. $M_{Schichtarbeit} > M_{keineSchichtarbeit}$) beurteilen zu können. Erst dann kann die H0 verworfen werden.*

*statistische Relevanz vs. klinische/inhaltliche Relevanz*

Bezogen auf das bereits erwähnte Beispiel (Qualität der Klient*innen-Therapeut*innen-Beziehung) würde ein signifikantes Ergebnis (p≤ 5%) bedeuten, dass sich die beiden Therapieformen (Therapieform A und Therapieform B) in der eingeschätzten Qualität der Klient*innen-Therapeut*innen Beziehung unterscheiden. Die Ausprägung des Signifikanzwerts (p-Wert) ist jedoch nicht nur von der Größe des beobachteten Unterschiedes, sondern ebenfalls von weiteren Faktoren wie etwa der Stichprobengröße abhängig (vgl. Jones, Carley & Harrison, 2003; Krzywinski & Altman, 2013). Eine ausschließliche Interpretation eines Ergebnisses anhand des Signifikanzwertes (p-Wert) greift somit in den meisten Anwendungsfällen zu kurz, da eine statistische Relevanz (signifikanter Unterschied) nicht mit einer inhaltlichen Relevanz (bedeutsamer Unterschied) gleichzusetzen ist. Insofern sollte jede statistisch signifikante Differenz hinsichtlich ihrer inhaltlichen Bedeutsamkeit beurteilt werden (zur Beurteilung z.B. klinisch relevanter Therapieeffekte siehe Kleist, 2010). Im angeführten Beispiel könnten die Therapieformen beispielsweise Einschätzungen im durchschnittlichen Bereich (z.B. in normierten T-Werten (unterdurchschnittlich: Werte <40; Durchschnittsbereich: Werte von 40-60; überdurchschnittlich: Werte >60): MTherapieA= 54, MTherapieB= 56) aufweisen (Szenario 1). In diesem Fall hätte die geringe Differenz von zwei T-Werten eine lediglich sehr geringe inhaltliche Relevanz, da beide Gruppen (Therapieform A und B) Mittelwerte im gleichen T-Wertebereich (durchschnittlich) aufweisen. Da der p-Wert nicht nur von der Größe des beobachteten Unterschiedes, sondern unter anderem auch von der Stichprobengröße abhängig ist, könnte die signifikante Differenz in diesem Szenario nur deshalb beobachtet worden sein, weil für die Berechnungen nicht die optimale, sondern eine zu große Personenanzahl (Oversampling) herangezogen wurde (vgl. Jones, Carley & Harrison, 2003; Krzywinski & Altman, 2013).

*Konfidenzintervalle als Hilfsmittel zur Bestimmung der Ergebnisrelevanz*

Zeigt sich hingegen ein Szenario (2), innerhalb dessen die beiden Therapieformen Mittelwerte in unterschiedlichen T-Wertebereichen aufweisen (z.B. Therapieform A mit M= 76 im überdurchschnittlichen und Therapieform B mit M= 50 im durchschnittlichen Bereich), kann dem Ergebnis eine größere inhaltliche Relevanz beigemessen werden. Dann wären Aussagen wie: „Die Therapieformen weisen unterschiedliche wahrgenommene Beziehungsqualitäten auf, wobei diese bei A eine überdurchschnittlich positive Ausprägung findet. Bei Therapieform B zeigt sich hingegen nur eine durchschnittliche Ausprägung der eingeschätzten Beziehungsqualität.“ möglich. Insgesamt muss jedoch beachtet werden, dass ein Populationsunterschied anhand der für die Berechnungen herangezogenen Stichproben geschätzt wird und diese Schätzung stets einer Zufallsschwankung unterliegt.

Um diese Unsicherheit in der Schätzung abzubilden, sollten Konfidenzintervalle herangezogen werden. Konfidenzintervalle spiegeln einen bestimmten Wertebereich wider, in dem der tatsächliche (unbekannte) Wert der Grundgesamtheit, mit einer vorab festgelegten Wahrscheinlichkeit (in der Regel 95%), zu verorten ist (vgl. Bender & Lange, 2007). Je kleiner der beobachtete Wertebereich ausfällt, desto „exakter" zeigt sich die auf der Basis der Stichprobe geschätzte Mittelwertdifferenz. Die Breite des Wertebereiches hängt dabei sowohl von der Größe als auch den Standardabweichungen der beiden Gruppen ab.

*Effektstärken als Hilfsmittel zur Bestimmung der Ergebnisrelevanz*

Zur weiteren Beurteilung der inhaltlichen Relevanz besteht, neben der Interpretation deskriptiver Kennwerte beider Gruppen (je nach angewandtem Verfahren z.B. Mittelwert (M), Standardabweichung (SD), Median (Mdn)), bei einigen Auswertungsmethoden die Möglichkeit sogenannte Effektstärken zu berechnen. Effektstärken sind ein Maß für die Größe eines Unterschiedes und stellen in der Regel standardisierte Kenngrößen dar, welche einen studienübergreifenden Ergebnisvergleich ermöglichen. Bezogen auf das Szenario (2) des Klient*innen-Therapeut*innen-Beziehung-Beispiels stellt sich die Frage, ob eine signifikante Mittelwertdifferenz von 26 T-Werten (MTherapieA= 76, MTherapieB= 50 oder umgekehrt) nun als klein oder groß zu beurteilen ist. Die Berechnung einer Effektstärke erlaubt die Größe des in der aktuellen Studie beobachteten Unterschiedes, mit Effekten aus ähnlichen Untersuchungen (z.B. Studien in anderen Ländern) zu vergleichen. Die Größe sowie inhaltliche Relevanz eines beobachteten Unterschiedes kann durch die Verwendung von Effektstärken, einfacher vor dem Hintergrund bisheriger Forschungsergebnisse interpretiert werden. Zeigt sich beispielsweise ein deutlich größerer/kleinerer Effekt als in bisherigen Forschungsarbeiten, eröffnen sich weitere Möglichkeiten ein Ergebnis zu diskutieren. Auch im Rahmen von Studien, in denen kleine Stichproben untersucht werden, lassen sich Effektstärken als essentieller Bestandteil der Ergebnisinterpretation anführen. Die Anwendung von inferenzstatistischen Unterschiedstestungen bei (zu) kleinen Stichproben geht in der Regel mit einem höheren Beta-Fehler-Risiko einher, d.h. weist eine höhere Wahrscheinlichkeit auf, eine vorhandene Systematik (Unterschied) nicht als solche identifizieren zu können. Effektstärken lassen sich in solchen Fällen auch als Indikator für eine zu geringe Stichprobengröße heranziehen, wodurch eine exaktere Interpretation des Ergebnisses ermöglicht wird. Lässt der p-Wert mit .070 beispielsweise auf einen nicht signifikanten Unterschied schließen, die Effektstärke weist jedoch auf einen mittleren Effekt hin, könnte das nicht signifikante Ergebnis lediglich auf eine zu geringe Stichprobengröße zurückzuführen sein (vgl. Jones, Carley & Harrison, 2003; Krzywinski & Altman, 2013). An dieser Stelle sei auf die Bedeutung einer eingehenden Stichprobenkalkulation zu verweisen, um eine möglichst hohe Testgenauigkeit zu erhalten (siehe Implikationen für die Praxis in diesem Beitrag).

*Effektstärken und (zu) kleine Stichproben*

*Interpretation von Effektstärken*

Als Bezugsrahmen für die Beurteilung der Stärke eines beobachteten Unterschiedes sollten stets Effekte aus vergleichbaren Untersuchungen herangezogen werden. Dennoch ermöglichen einige Effektstärken, beispielsweise Cohens d, die Beurteilung der Größe eines Unterschiedes ohne den Einbezug bisheriger Forschungsergebnisse anhand festgelegter

Effektintervalle (siehe Exkurs). Bezugnehmend auf die drei erwähnten Verfahren (t-Test, Welch-Test, U-Test) verfügen lediglich der t-Test sowie der Welch-Test über eine gängige Effektstärke (Cohens d). Im Zuge der Anwendung eines Mann-Whitney-U-Tests ist es hingegen unüblicher eine Effektgröße anzuführen, dennoch sollte anhand des über den U-Test erhaltenen z-Werts die Effektstärke r berechnet werden (vgl. Fritz, Morris & Richler, 2012).

***Exkurs: Effektstärke Cohens d***

*Die Effektstärke Cohens d wird anhand der Mittelwerte der beiden zu vergleichenden Gruppen sowie der gepolten Standardabweichung berechnet (Cohen, 1988). Da die Mittelwerte der zu vergleichenden Gruppen die Basis für die Berechnung darstellen, sollte diese Effektstärke nur bei normalverteilten Merkmalen herangezogen werden. Um die Interpretation der Effektstärke zu vereinfachen, beschreibt Cohen (1988) drei Intervalle, welche verschiedene Effektkategorien widerspiegeln. So handelt es sich ab einem Wert von d= ±.20 um einen kleinen, ab d=±.50 um einen mittleren und ab d=±.80 um einen großen Effekt. Diese Intervalle dienen jedoch in erster Linie einer einfacheren Interpretation des Ergebnisses und sollten nicht als allgemein gültige Richtlinien betrachtet werden.*

$$d = \frac{M_1 - M_2}{S_{gesamt}}$$

$$S_{gesamt} = \frac{(N_1 * S_1) + (N_2 * S_2)}{N_1 + N_2}$$

*Skalenniveau des zu untersuchenden Merkmals (abhängige Variable) als Kriterium für die Verfahrenswahl*

Nachdem die Gemeinsamkeiten in der Berechnung sowie der Interpretation der Ergebnisse der drei Verfahren erläutert wurden, bleibt noch zu klären, in welchen spezifischen Ausgangssituationen, welches der Verfahren Anwendung findet. Wie bereits erwähnt, hängt die Verfahrenswahl vor allem von den vorhandenen Forschungsdaten ab. In erster Linie bestimmt das Skalenniveau des zu untersuchenden Merkmals (abhängige Variable, hinsichtlich wessen? z.B. Therapeut*innen-Klient*innen-Beziehung) die Auswahl des korrekten Auswertungsverfahrens. Anhand der Frage „Wen vergleiche ich hinsichtlich wessen?“, lässt sich feststellen, welche Variable als Gruppierungsfaktor dient (Wen?, unabhängige Variable) und welches Merkmal (hinsichtlich wessen?, abhängige Variable) untersucht werden soll. Der t-Test sowie der Welch-Test lassen sich ausschließlich zur Untersuchung metrischer (verhältnis- oder intervallskalierter) Variablen heranziehen, wohingegen der Mann- Whitney-U-Test sowohl bei metrischen als auch bei ordinalen Variablen angewandt werden kann. Soll beispielsweise untersucht werden, ob sich ein signifikanter Unterschied in der Einschätzung der Zufriedenheit mit dem angebotenen Therapieprogramm einer Tagesklinik (5= sehr zufrieden bis 1= überhaupt nicht zufrieden) zwischen weiblichen und männlichen Patient*innen beobachten lässt, kann aufgrund des

*t-Test als Verfahren der Wahl bei metrisch skalierten Merkmalen*

Skalenniveaus der untersuchten Variable (Zufriedenheitseinschätzung = ordinales Skalenniveau) lediglich der Mann-Whitney-U-Test herangezogen werden. Weist die untersuchte Variable hingegen ein metrisches Skalenniveau auf (z.B. Messung der Zufriedenheit mit einem standardisierten Fragebogen), kann sowohl der t-Test für unverbundene Stichproben als auch der Mann-Whitney-U-Test angewandt werden. Dennoch sollte bei der Untersuchung von metrischen Variablen, wenn möglich, auf den t-Test zurückgegriffen werden, da dieser gegenüber dem U-Test das mächtigere Verfahren darstellt, d.h. über eine höhere Wahrscheinlichkeit verfügt, eine vorhandene Systematik (Unterschied) korrekterweise zu identifizieren. Neben der Bestimmung des Skalenniveaus des untersuchten Merkmals (Zielvariable hinsichtlich derer die Gruppen verglichen werden) müssen vor allem beim t-Test für unverbundene Stichproben verfahrensspezifische Voraussetzungen (z.B. Normalverteilung des untersuchten Merkmals sowie Varianzhomogenität in beiden Stichproben) erfüllt sein, um nach einem traditionellen Vorgehen eine korrekte Anwendung der Auswertungsmethode zu gewährleisten.

## *Wahl des korrekten Verfahrens*

Wie im vorangegangenen Abschnitt bereits behandelt wurde, ist die Auswahl des korrekten Verfahrens neben dem Messniveau der Variable, hinsichtlich derer zwei unverbundene Stichproben verglichen werden, ebenso abhängig von unterschiedlichen empirischen Voraussetzungen. Folgend werden die drei in diesem Artikel, vorgestellten Verfahren – t-Test, Welch- und U-Test – bezüglich derer mathematischen Funktionsweise sowie spezifischen Voraussetzungen besprochen.

### *t-Test für zwei unverbundene Stichproben:*

Der t-Test für unverbundene Stichproben – auch Zwei-Stichproben-t-Test genannt – stellt von den drei Verfahren dasjenige mit der höchsten statistischen Testmacht dar. Allerdings verlangt dieses Verfahren auch die Einhaltung der meisten empirischen Voraussetzungen, auf deren Prüfung vor Anwendung dieses Verfahrens besonders Acht gegeben werden sollte.

*Prüfung der Normalverteiluna*

Zunächst muss die Variable bzw. das Merkmal, hinsichtlich dessen die zwei unverbundenen Stichproben verglichen werden, mindestens intervall-skaliert sein. Dieses mindestens intervall-skalierte Merkmal muss in beiden Stichproben normalverteilt sein, was die erste zu überprüfende Voraussetzung im Rahmen dieses Testverfahrens darstellt. Um diese Voraussetzung zu überprüfen, stehen Forscherinnen und Forschern mehrere Vorgehensweisen zur Verfügung:

Besonders bei kleineren Stichproben ($n < 30$) ist eine Prüfung der Normalverteilung sinnvoll. Im Falle einer größeren Stichprobe ($n \geq 30$) kann diese jedoch gemäß des Zentralen Grenzwertsatzes vernachlässigt werden. Daher wird bei einer Ausgangslage von Gruppen bzw. Stichprobengrößen von $n \geq 30$ häufig auf eine weitere Überprüfung der Normalverteilung verzichtet.

*Schiefe und Kurtosis zur Prüfung der NV*

Zur Überprüfung der Normalverteilung in jeder Gruppe besteht die Möglichkeit, entlang deskriptiv-statistischer Kennwerte über das Vorliegen der Verteilungsanforderung zu befinden. Dabei können vor allem zwei unterschiedliche Kennwerte herangezogen werden: Schiefe und Kurtosis. Die Schiefe zeigt, ob und wie stark eine Verteilung nach links bzw. nach rechts geneigt ist. Liegen Hinweise auf eine Normalverteilung vor, so nimmt die Verteilung einen symmetrischen Verlauf an und weist Werte zwischen -1 und + 1 auf. Die Kurtosis zeigt an, wie steil bzw. gewölbt die Werte einer Merkmalsausprägung verteilt sind. Um auf normalverteilte Daten schließen zu können, sollten diese ebenfalls Werte zwischen -1 und + 1 annehmen.

*Histogramm zur Prüfung der NV*

Zudem gibt es eine graphische Option der Überprüfung der Normalverteilung, nämlich das sogenannte Histogramm. Dabei werden eine Normalverteilungskurve angezeigt und die zusammengefassten Messwerte in Form von Balken dargestellt. Die dargestellten Balken sollten, um auf das Vorliegen einer Normalverteilung schließen zu können, einen möglichst symmetrischen Verlauf abbilden (s. Abb. 1), wobei grobe Abweichungen als ein Indikator für keine Normalverteilung angesehen werden können/dürfen (s. Abb.2).

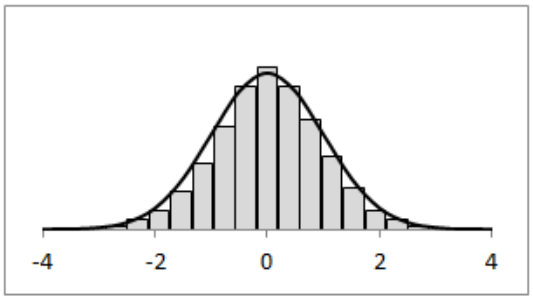



Abb.1: Histogramm (Normalverteilung)

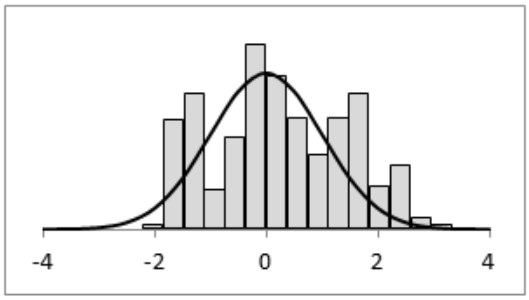



Abb.2: Histogramm (keine Normalverteilung)

*Anpassungstests zur Prüfung der NV*

Des Weiteren kann die Voraussetzung der Normalverteilung auch mit Hilfe statistischer Tests überprüft werden. Hierbei wird der Test nach Kolmogorov-Smirnov und/oder Shapiro-Wilk angewendet. Diese überprüfen, ob ein oder mehrere Merkmal(e) normalverteilt ist/sind. Bei der Interpretation dieses Verfahrens ist der Signifikanzwert (p-Wert) ausschlaggebend, welcher im Falle einer Normalverteilung keine Signifikanz ($p > .05$) anzeigen darf (Bortz, 2016).

Allerdings reagiert der t-Test auf eine Verletzung dieser Voraussetzung robust, sodass mit keinen gröberen Fehlschlüssen bei der Ergebnisinterpretation zu rechnen ist (z.B. Rasch, Kubinger & Moder, 2011).

*Vorsicht bei Ausreißern*

Vorsicht ist jedoch besonders bei kleinen Stichproben geboten, welche häufiger in der klinischen Forschung auftreten. Hier können sogenannte „Ausreißer“ (extrem hohe bzw. extrem niedrige Messwerte einzelner Personen, die den Mittelwert verzerren) zu Fehlinterpretationen führen. Obwohl es sich bei der Frage, ob Ausreißer vorliegen, um keine Voraussetzung im eigentlichen Sinne handelt, ist dieser Umstand doch stets und vor allem bei kleinen Stichproben vorab zu untersuchen. Eine Prüfung erfolgt in der Regel

graphisch über die Betrachtung von sogenannten Boxplots.

*Homogenität der Varianzen*

Die zweite empirisch prüfbare Voraussetzung für die Anwendung des t-Tests für unverbundene Stichproben stellt die Homogenität der Varianzen dar. Die Voraussetzung der Varianzhomogenität bedeutet, dass die beiden Stichproben im untersuchten Merkmal eine ähnliche Streuung aufweisen sollten, d.h. die Variation des Merkmals, die innerhalb der einen Gruppe besteht, mit der Variation innerhalb der anderen Gruppe in einem vergleichbaren Ausmaß vorliegt. Zu überprüfen ist diese Voraussetzung beispielsweise mit dem sogenannten Levene-Test. Wenn dieser Test einen p-Wert von >.05 aufweist, d.h. nicht signifikant ausfällt, kann diese Voraussetzung als gegeben angesehen werden.

*Verletzung der Voraussetzungen*

Auf eine Verletzung dieser Voraussetzung reagiert der t-Test – ähnlich wie gegenüber einer Verletzung der Normalverteilung – robust (z.B. Sawilowsky & Blair, 1992; Rasch, Kubinger & Moder, 2011). Allerdings ist eine Verletzung der Varianzhomogenität bei bestimmten empirischen Ausgangssituationen sehr wohl mit Folgeproblemen für eine korrekte Interpretation der Ergebnisse verbunden. In einem Szenario, bei dem sowohl die Stichprobengrößen der beiden zu vergleichenden Gruppen als auch deren Varianzen ungleich sind, besteht die Gefahr statistischer Fehlschlüsse im Sinne eines Alpha- oder Beta-Fehlers. Weist die Gruppe mit der geringeren Personenzahl breiter gestreute Messwerte auf (höhere Varianz) als die Gruppe mit der höheren Personenzahl, erhöht sich die Wahrscheinlichkeit einen scheinbar systematischen Unterschied zu identifizieren, welcher jedoch so in der Population nicht vorhanden ist (Alpha-Fehler) (z.B. Ramsey, 1980). Im umgekehrten Fall, dass die Gruppe mit der höheren Personenzahl auch eine größere Varianz aufweist als die Gruppe mit der geringeren Personenzahl, fällt die Testentscheidung eher konservativer aus, was mit einem höheren Risiko einhergeht, einen systematischen Unterschied zu negieren, obgleich dieser in der Population vorhanden ist (Beta-Fehler).

Bei einem „ traditionellen“ Vorgehen zur korrekten Verfahrenswahl wird jedoch grundsätzlich beim Vorliegen heterogener Varianzen auf das Verfahren des Welch-Tests ausgewichen, der in einem späteren Absatz dieses Beitrags näher behandelt wird.

Gemäß der traditionellen Entscheidungskriterien darf folglich erst bei Erfüllung aller Voraussetzungen (Normalverteilung des Merkmals in beiden Gruppen, keine Ausreißer bei kleinen Stichproben, Homogenität der Varianzen) der t-Test für unverbundene Stichproben eingesetzt werden, der folgend näher betrachtet werden soll.

*Berechnung des empirischen t-Werts*

Der t-Test für unverbundene Stichproben erhält seinen Namen von der stochastischen Prüfverteilung, derer sich zur Beurteilung einer empirisch ermittelten Mittelwertdifferenz bedient wird. In der Formel wird dies beim Blick auf das forcierte Ergebnis, nämlich einen empirischen Verteilungswert t, deutlich.

$$t = \frac{\overline{X_1} - \bar{X}_2}{\sigma_0 * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$\sigma_0 = \frac{(n_1 - 1) * \widehat{\sigma_1^2} + (n_2 - 1) * \widehat{\sigma_2^2}}{n_1 + n_2 - 2}$$

Dieser t-Wert stellt eine standardisierte Differenz derjenigen Mittelwerte dar, die von beiden Gruppen im zu untersuchenden Merkmal erreicht wird. Die Standardisierung erfolgt durch die gepolte Standardabweichung σ_(0.  ) σ_0  stellt seinerseits eine Form von gemittelter Standardabweichung dar, die sich aus Mittelung/Polung der standardisierten Streuung beider Stichproben ergibt. Die so ermittelte und in Gestalt des t-Wertes repräsentierte Mittelwertdifferenz zweier Gruppen bildet in weiterer Folge die Grundlage für die Verallgemeinerung des berechneten Ergebnisses, also ob von dieser auf Basis der vorliegenden Stichprobe ermittelten Differenz auf die Population geschlossen werden kann, sprich, das Ergebnis signifikant ist. Der errechnete t-Wert allein reicht jedoch noch nicht aus, um einen solchen Schluss ziehen zu können. Zur Beurteilung der Signifikanz werden zusätzlich die sogenannten Degrees of Freedom (Df) benötigt. Diese ergeben sich im Wesentlichen aus der Stichprobengröße und werden wie folgt berechnet:

*Freiheitsgrade (df)*

$$df = n_1 + n_2 - 2$$

*Reportkonventionen (APA) t-Test für unverbundene Stichproben*

Die American Psychological Association schreibt in ihren Richtlinien (APA, 2020) vor, dass Resultate eines t-Tests unter Angabe spezifischer Kennwerte erfolgen sollten, um das Ergebnis transparent darstellen und in weiterer Folge dessen Bewertung besser nachvollziehen zu können. Konkret sollte der statistische Report des Ergebnisses (1) den empirischen t-Wert, (2) die *Degrees of Freedom*, (3) den p-Wert, (4) die Effektstärke d, (5) die Mittelwerte/Standardabweichungen der beiden Gruppen sowie optional (6) das Konfidenzintervall der Mittelwertdifferenz enthalten.

Angenommen es sei beim Vergleich zweier unverbundener Stichproben ein t-Wert von 3.74, bei df von 974, mit einem p-Wert von p< .001, $M_{Gruppe1}$= 54.00 (SD= 10.22), $M_{Gruppe2}$= 51.57 (SD= 10.05), wobei die Effektstärke  d= .38 beträgt, gegeben. Gemäß der Reportkonvention wären diese Kennwerte nach Testung der Hypothesen wie folgt anzuführen:

*(t(974)= 3.74, p< .001, d= 0.38, MG1= 54.00 (SD= 10.22), MG2= 51.57 (SD= 10.05), 95% CI [1.16, 3.72])*

### *Welch-Test*

Der Welch-Test entstand als Antwort auf das sogenannte *Behrens-Fisher-Problem*, welches beim Vorliegen ungleicher Varianzen beim Mittelwertvergleich zweier unverbundener Stichproben entsteht. Neben anderen Forschern schlug auch Bernard Welch (1947) einen

*Welch-Test beim Vorliegen heterogener Varianzen*

approximativen Ansatz als möglichen Umgang mit diesem Problem vor (für vertiefende Informationen siehe im Original: Welch, 1947). Dieser Ansatz findet sich im Welch-Test wieder, dessen Einsatz beim Vorliegen heterogener Varianzen indiziert ist. Bei diesem Testverfahren wird der Standardfehler der Mittelwertdifferenz geschätzt und, unter Berücksichtigung dessen, ein *t*-Wert berechnet.

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{\widehat{\sigma_1^2}}{n_1} + \frac{\widehat{\sigma_2^2}}{n_2}}}$$

Die diesem t-Wert zugehörigen Degrees of Freedom (Df) werden unter Berücksichtigung des Faktors c wie folgt berechnet:

$$df = \frac{(n_1 - 1) * (n_2 - 1)}{(n_2 - 1) * c^2 + (n_1 - 1) * (1 - c)^2}$$

$$c = \frac{\frac{\widehat{\sigma_1^2}}{n_1}}{\frac{\widehat{\sigma_1^2}}{n_1} + \frac{\widehat{\sigma_2^2}}{n_2}}$$

Der Welch-Test weist gegenüber dem t-Test eine geringere Testmacht auf, weshalb bei Vorliegen der Voraussetzungen dem t-Test Vorzug zu geben ist.

*Reportkonventionen Welch-Test*

Angenommen es ergebe sich beim Vergleich zweier unverbundener Stichproben beispielsweise ein t-Wert von 2.38, bei df= 90.08, mit einem p-Wert von p= .020, $M_{Gruppe1}$= 15.23 (SD= 3.88), $M_{Gruppe2}$= 13.31 (SD= 3.91), wobei die Effektstärke d= 0.49 beträgt. Der Welch-Test folgt denselben Reportkonventionen wie der t-Test, woraus sich Folgendes ergibt:

*(t(90.08)= 2.38, p= .020, d= 0.49, MG1= 15.23 (SD= 3.88), MG2= 13.31 (SD= 3.91), 95% CI [0.32, 3.53])*

### *U-Test*

*Entscheidung für den U-Test*

Der U-Test von Mann und Whitney (1947) stellt ein Verfahren der sogenannten verteilungsfreien bzw. non-parametrischen Tests dar. Dessen Anwendung ist dabei nicht wie beim t- oder Welch-Test an eine spezifische Verteilungsanforderung, wie jene der Normalverteilung gebunden. Der Mann-Whitney-U-Test kann bei zwei empirischen Ausgangssituationen als Verfahren der Wahl angesehen werden. Zum einen ist dieses Verfahren – einer bestimmten methodischen Tradition folgend – indiziert, wenn zwei unabhängige Stichproben hinsichtlich eines mindestens intervall-skalierten Merkmals verglichen werden sollen, die Voraussetzung der Normalverteilung des Merkmals aber nicht erfüllt ist. In solchen Fällen wird bei einem traditionellen Vorgehen üblicherweise der U-Test herangezogen, da dieser auch im Vergleich zu anderen non-parametrischen Alternativen, wie dem Medientest, eine höhere Teststärke/Testmacht aufweist (Bortz & Lienert, 2008).

Zum anderen kann dieser eingesetzt werden, wenn zwei unverbundene Stichproben hinsichtlich eines rang- bzw. ordinal-skalierten Merkmals verglichen werden sollen. Der U-Test fungiert somit sowohl als Ausweichverfahren bei nicht erfüllten Verteilungsanforderung des t- oder Welch-Tests als auch als eigenständiges, non-parametrisches Verfahren zur Testung unterschiedlicher Medianausprägungen einer Population. Gegenüber dem t-Test ist der U-Test jedoch mit einer geringeren Testmacht ausgestattet, weshalb der t-Test bei Erfüllung aller Voraussetzung vorzuziehen ist.

*U-Test und Rangtransformation*

Im Vergleich zum t- und Welch-Test wird im Rahmen des U-Tests nicht mit den tatsächlichen Messwerten gerechnet, sondern diese mittels Rangtransformation in Ränge überführt. Hierfür wird pro Gruppe jedem Individuum ein Rang, entsprechend der Ausprägung seines Messwertes, zugeordnet (kleinster Wert Rang 1 bis größter Wert Rang n). Anschließend werden pro Gruppe Rangsummen gebildet, anhand derer die U-Werte zur Bestimmung der Signifikanz berechnet werden. Bei ausreichend großen Stichproben (N > 20) lässt sich die Prüfgröße U unter Verwendung einer Normalverteilungsapproximation in die Standardnormalverteilung (z-Verteilung) überführen (vgl. Bortz & Lienert, 2008). Die Bestimmung der Signifikanz erfolgt in solchen Fällen anhand einer Z-Prüfgröße.

*Reportkonventionen U-Test*

Auch im Rahmen des U-Tests, sollte die Ergebnisdarstellung bestimmte Kennwerte enthalten. Konkret handelt es sich hierbei um (1) den empirischen U-Wert, (2) den empirischen z-Wert bei ausreichend großen Stichproben (N > 20) sowie (3) den p-Wert. Zusätzlich sollten, zur besseren Interpretierbarkeit des Ergebnisses, die Mediane beider Gruppen, sowie ein Maß für die Effektstärke (z.B. r, Fritz, Morris & Richler, 2012) angeführt werden. Die Effektstärke r lässt sich mit folgender Formel berechnen:

*Effektstärke r im Rahmen des U-Tests*

$$r = \frac{z}{\sqrt{N}}$$

Ergibt sich beim Vergleich zweier unverbundener Stichproben beispielsweise ein U-Wert von 411, ein z-Wert von 5.17 sowie ein p-Wert von <.001 mit einer Effektstärke von r= .54, sollten die Kennwerte wie folgt angegeben werden:

(*U= 411, Z= 5.17, p< .001, r= .54*)

## *Konklusion und Hinweise für die Forschungspraxis bei psychotherapiewissenschaftlichen Fragestellungen*

In der psychotherapiewissenschaftlichen Forschungspraxis sind Forscher*innen und Forscher immer wieder mit Frage- und Problemstellungen konfrontiert, die sich mit dem potentiellen Unterschied zwischen zwei unverbundenen Personengruppen befassen. Sei es, um geschlechtsspezifische Differenzen in einem psychische Morbidität anzeigenden Merkmal zu identifizieren oder auch Versuchs- und Kontrollgruppen in der Ausprägung einer evaluativen Zielvariable zur Baseline/vor einer Therapie zu vergleichen. Auch Unterschiede zwischen zwei Therapiemethoden, die nicht auf einen Wirksamkeitsvergleich abzielen, sondern vielmehr strukturelle Aspekte wie Anzahl der Sitzungen etc. adressieren, sind aus statistischer Perspektive ein Vergleich zweier unverbundener Stichproben bzw. Personengruppen.

Doch mit welchen konkreten methodischen Herausforderungen sind insbesondere Forscher*innen und Forscher auf dem Gebiet der Psychotherapiewissenschaft bei der Anwendung der drei vorgestellten Auswertungsverfahren, aus statistischer Perspektive, konfrontiert?

*Zuverlässigkeit von statistischen Entscheidungen als Hauptproblemfeld*

Das Hauptproblemfeld stellt die Testsicherheit bzw. die Zuverlässigkeit der statistischen Entscheidungen bezüglich des Verwerfens der H0 dar, die Frage, ob der auf Basis der vorliegenden Stichproben identifizierte Effekt als gesichert angenommen werden kann. In diesem Zusammenhang spielen der Fehler erster (α-Fehler) und der Fehler zweiter Art (β-Fehler) sowie die Macht (power) die entscheidende Rolle. Die oben genannten drei Aspekte hängen dabei von zahlreichen Faktoren ab, die mit einer sorgfältigen Studienplanung durch die Forscher*innen selbst besser kontrolliert werden können. Wie dies umgesetzt werden kann, wird in einem späteren Abschnitt näher thematisiert.

*Stichprobengröße als Problemfaktor(Over- bzw. Undersampling)*

In der Psychotherapieforschung sind Wissenschaftler*innen – besonders in der klinischen Forschung – mit geringen Fallzahlen konfrontiert. Die Berechnungen müssen dann an kleinen Stichproben erfolgen, was mit einigen statistischen Problemen verbunden ist. Einerseits wird durch eine kleine Stichprobe die Testmacht reduziert, was das Risiko eine vorhandene Systematik zu „übersehen“ (β-Fehler) erhöht. Bei Interpretation des Ergebnisses eines t-Tests bei unverbundenen Stichproben ist dem Umstand geringer Stichprobenumfänge Rechnung zu tragen, um statistische Fehlschlüsse zu vermeiden. Den p-Wert als alleinige Maßzahl zur Beurteilung eines Effekts heranzuziehen, ist insbesondere

bei nicht optimaler Stichprobengröße unzureichend, weshalb eine Betrachtung der Effektstärke d dringend zu empfehlen ist. Bei sehr großen Stichproben, die etwa im Rahmen groß angelegter Online-Befragungen generiert werden, ergibt sich bei alleiniger Betrachtung des p-Werts ein hohes Risiko in der statistischen Beurteilung des Ergebnisses einem α-Fehler zu unterliegen. In solchen Ausgangssituationen muss ein signifikantes Ergebnis ($p \leq .05$) keineswegs eine tatsächliche Systematik wiederspiegeln. Im Falle eines sog. Oversamplings können bereits kleinste Differenzen zu einem signifikanten Ergebnis führen. Die Interpretation des p-Werts muss also auch in einem solchen Fall unbedingt mit der Betrachtung der Effektstärke (beim t-Test z.B. d) einhergehen.

Die vorangegangenen Ausführungen machen deutlich, dass der p-Wert als alleinige Kennzahl (bei t-, Welch- und U-Test) für eine zuverlässige statistische Entscheidung nicht ausreicht!

➔ **Empfehlung 1: Interpretation eines Ergebnisses anhand des p-Werts sowie einer Effektstärke**

*Stichprobenkalkulation und Testmacht*

Um sowohl Over- als auch Undersampling, wie in den oben beschriebenen Szenarien, zu vermeiden und ein adäquates Verhältnis von Macht und Fehlerwahrscheinlichkeit sicherzustellen, sollte im Rahmen einer eingehenden Untersuchungsplanung der optimale Stichprobenumfang vorab kalkuliert werden (vgl. Chow, Shao, Wang & Lokhnygina, 2018), was mit frei zugänglichen Auswertungsprogrammen wie beispielsweise G*power oder R möglich ist. Forscherinnen und Forscher haben unter anderem damit die Möglichkeit, auf die Zuverlässigkeit des statistischen Ergebnisses Einfluss zu nehmen.

➔ **Empfehlung 2: Stichprobenkalkulation im Rahmen der Studienplanung**

*Verfahrenswahl und Testmacht*

Außerdem begünstigt die Wahl des statistischen Testverfahrens die Zuverlässigkeit weiter. So weist beispielsweise der t-Test, bei gegebenen Voraussetzungen, eine höhere Testmacht als der U-Test auf. Ein korrektes und reflektiertes Vorgehen in der Verfahrenswahl (siehe Anhang) stellt somit eine weitere Möglichkeit für Forscherinnen und Forscher dar, die Zuverlässigkeit des Ergebnisses aktiv zu beeinflussen.

➔ **Empfehlung 3: Wahl des korrekten statistischen Testverfahrens**

*statistische Relevanz ≠ inhaltliche Relevanz*

Neben der Frage nach der statistischen Zuverlässigkeit eines Ergebnisses spielt die Bestimmung der Ergebnisrelevanz eine zentrale Rolle in quantitativen Forschungsprozessen. Die Beobachtung eines statistisch signifikanten Ergebnisses ($p \leq .050$) bedeutet nicht, dass diesem Ergebnis ebenfalls eine inhaltliche Bedeutsamkeit beizumessen ist. Die Beurteilung eines Ergebnisses auf beiden Ebenen, statistisch sowie inhaltlich, schützt vor falschen Rückschlüssen und unsachgemäßen Implikationen für die Praxis. Auch im Rahmen psychotherapiewissenschaftlicher Untersuchungen sollte stets die Skalierung der jeweiligen Zielvariable (z.B. Normwerte: T-Skalierung) und die damit verbundenen bedeutungstragenden Intervalle bzw. Grenzwerte (z.B. T-Skalierung: unterdurchschnittlich: Werte <40; Durchschnittsbereich: Werte von 40-60; überdurchschnittlich: Werte >60) in die

Ergebnisinterpretation miteinbezogen werden. Dieser Aspekt ist vor allem bei Effektivitätsnachweisen von Bedeutung.

➔ **Empfehlung 4: inhaltliche vs. statistische Relevanz reflektieren**

Angesichts methodischer Kritik, die an vielen psychotherapiewissenschaftlichen Studien geübt wird, ist ein exaktes Vorgehen unter Einhaltung der methodischen Konventionen essentiell für die Glaubwürdigkeit eigener Forschungsergebnisse. Ein Bewusstsein für die geschilderten Sachverhalte sowie die Berücksichtigung der angeführten Empfehlungen, vermögen zu einer Steigerung der Ergebnisqualität psychotherapiewissenschaftlicher Forschungen mit unverbundenen Stichproben beizutragen, die argumentative Ausgangsposition zu stärken und darüber hinaus die Anfechtbarkeit der Ergebnisse zu reduzieren.

## *Literaturverzeichnis*

American Psychological Association. (2020). *Publication Manual of the American Psychological Association* (7th. Ed.). APA: Washington, DC.

Berth, H. & Brähler, E. (2003). Bonner Fragebogen für Therapie und Beratung - Testinformation. *Diagnostica, 94* (4). 191-194.

Bortz, J. (2006). *Statistik: Für Human-und Sozialwissenschaftler*. Springer Medizin Verlag: Heidelberg.

Bortz, J., & Lienert, G. A. (2008). *Kurzgefasste Statistik für die klinische Forschung: Leitfaden für die verteilungsfreie Analyse kleiner Stichproben*. Springer-Verlag.

Chow, S.C., Shao, J., Wang, H., Lokhnygina, Y. (2018). *Sample Size Calculations in Clinical Research*. New York: Chapman and Hall/CRC.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*, 2nd ed. Hillsdale, NJ: Erlbaum.

Fritz, O. F., Morris, P. E. & Richer, J. J. (2012). Effect Size Estimates: Current Use, Calculations, and Interpretation. *Journal of Experimental Psychology, 141* (1), 2–18.

Hagemann, W. & Geuenich, K. (2009). *Burnout-Screening-Skalen (BOSS)*. Göttingen: Hogrefe.

Jones, S. R., Carley, S. & Harrison, M. (2003). An introduction to power and sample size estimation. *Emergency Medicine Journal, 20*, 453-458.

Kleist, P. (2010). Wann ist ein Studienergebnis klinisch relevant?. Swiss Medical Forum, 10 (32), 525-527.

Krzywinski, M. & Altman, N. (2013). Power and sample size. *Nature Methods, 10*, 1139-1140.

Kühner, C., Bürger, C., Keller, F. & Hautzinger, M. (2007). Reliabilität und Validität des revidierten Beck-Depressionsinventars (BDI-II). Befunde aus deutschsprachigen Stichproben. *Der Nervenarzt, 78*, 651-656.

Mann, H. B., & Whitney, D. R. (1947). *On a test of whether one of two random variables is stochastically larger than the other. The annals of mathematical statistics*, 50-60.

Rasch, D., Kubinger, K. D., & Moder, K. (2011). The two-sample t test: pre-testing its assumptions does not pay off. *Statistical papers, 52* (1), 219-231.

Ramsey, P. H. (1980). Exact type 1 error rates for robustness of student's t test with unequal variances. *Journal of Educational Statistics, 5* (4), 337-349.

Sawilowsky, S. S., & Blair, R. C. (1992). A more realistic look at the robustness and type II error properties of the t test to departures from population normality. *Psychological bulletin, 111* (2), 352.

Welch, B. L. (1947). The generalization of student's' problem when several different population variances are involved. *Biometrika, 34* (1/2), 28-35.

## Autoren

Institut für Statistik

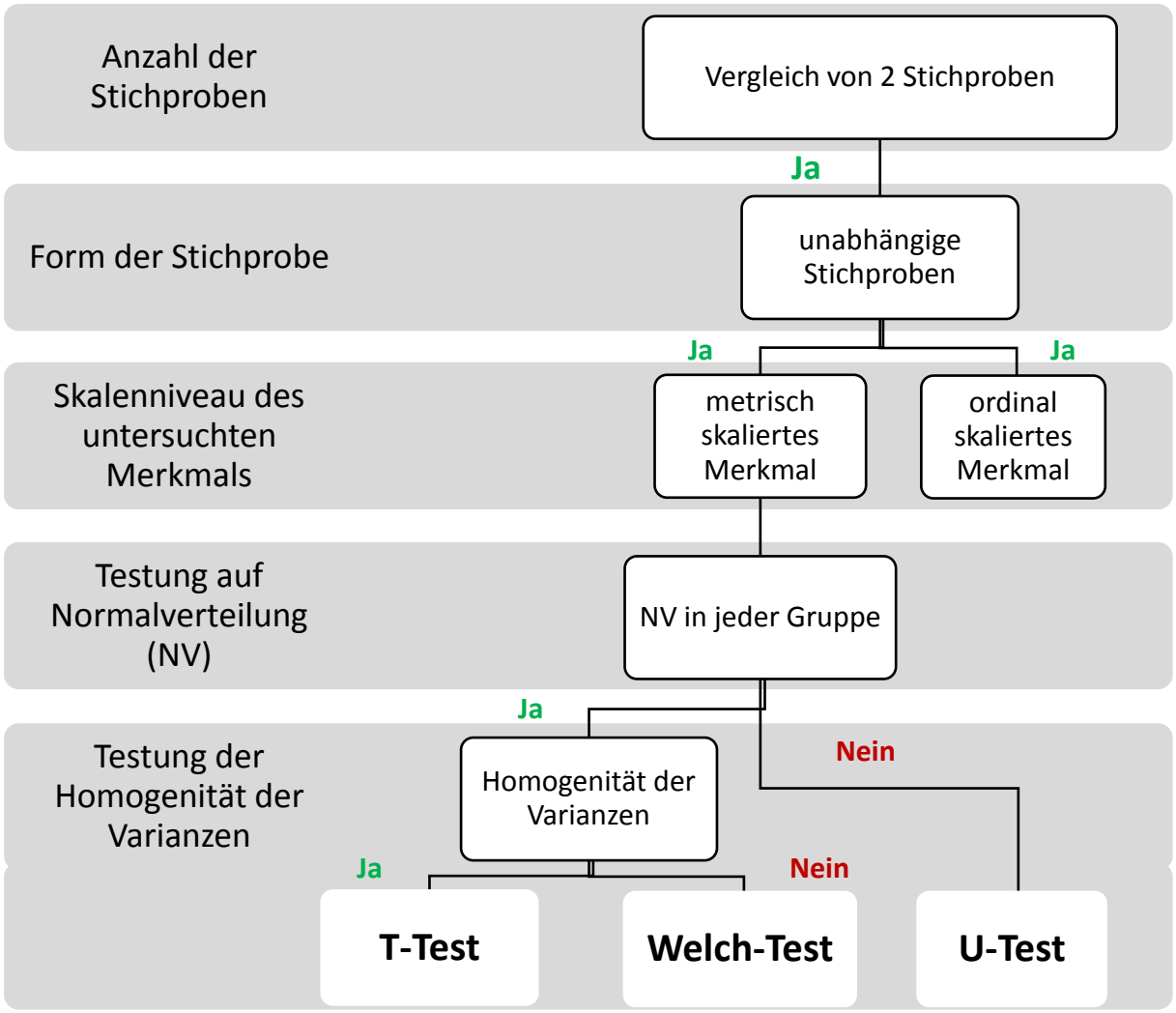
David Seistock, Anastasiya Bunina, Jan Aden

Adresse: Freudplatz 1, 1020 Wien

Raum 6011

E-Mail: jan.aden@sfu.ac.at

## *Anhang*


Anzahl der Stichproben
Vergleich von 2 Stichproben
Ja
Form der Stichprobe
unabhängige Stichproben
Ja
Ja
Skalenniveau des untersuchten Merkmals
metrisch skaliertes Merkmal
ordinal skaliertes Merkmal
Testung auf Normalverteilung (NV)
NV in jeder Gruppe
Ja
Nein
Testung der Homogenität der Varianzen
Homogenität der Varianzen
Ja
Nein
T-Test
Welch-Test
U-Test