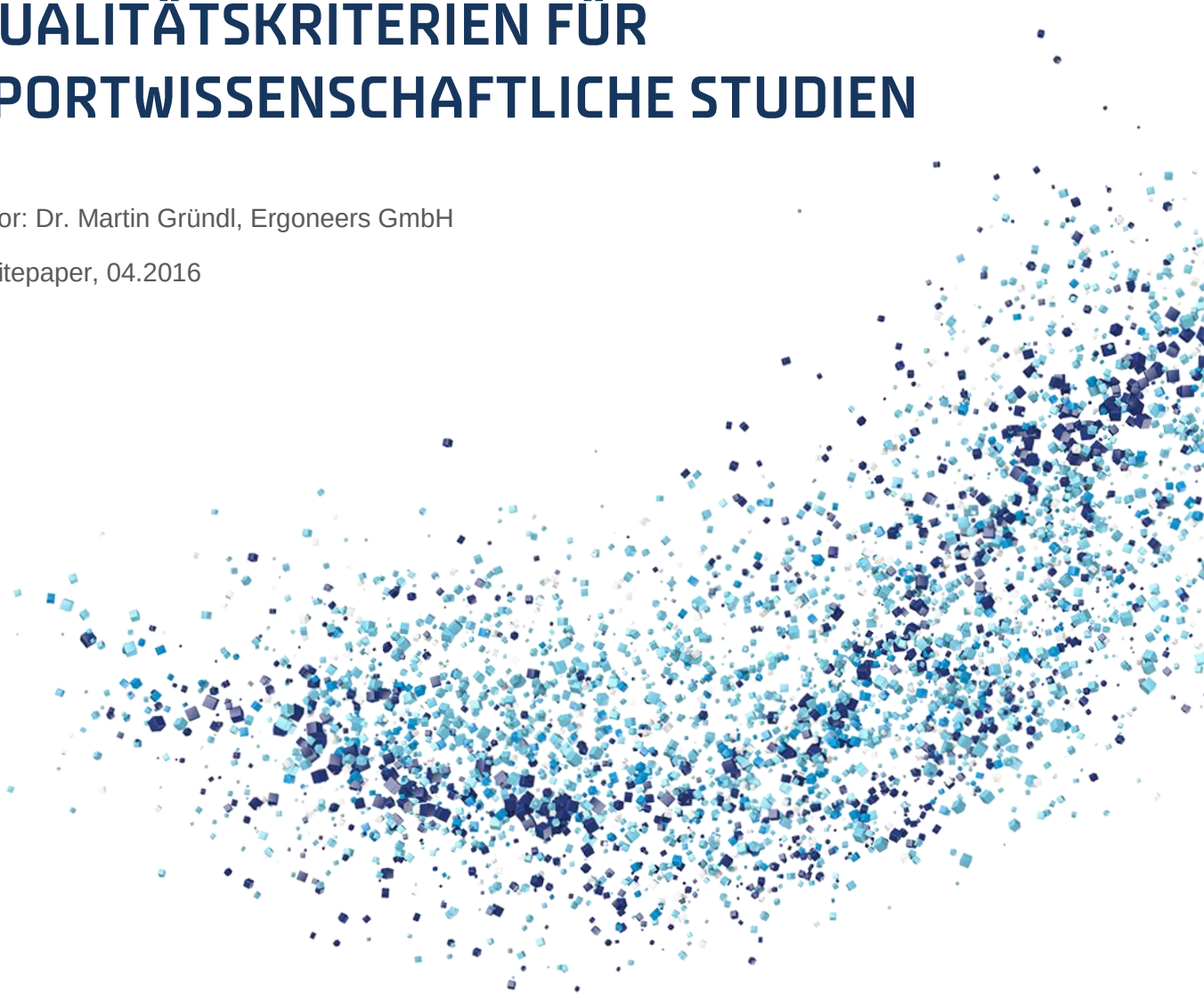
WISSENSCHAFTLICHE QUALITÄTSKRITERIEN FÜR SPORTWISSENSCHAFTLICHE STUDIEN

Autor: Dr. Martin Gründl, Ergoneers GmbH

Whitepaper, 04.2016



ERGONEERS
FROM SCIENCE TO INNOVATION

ABSTRACT

Die Verbesserung von Trainingsmethoden ist eine zentrale Aufgabe der Sportwissenschaft. Dazu gehört auch die Evaluation dieser Trainingsmethoden, für die Sportwissenschaftler Erkenntnisse und Methoden mehrerer anderer wissenschaftlicher Disziplinen wie der Naturwissenschaften, Sozialwissenschaften und der Sportmedizin nutzen. Es werden beispielsweise Experimente im Vorher-Nachher-Design durchgeführt, Experimentalgruppen mit Kontrollgruppen verglichen, es werden leistungsdiagnostische Tests angewandt, physiologische Parameter erhoben und am Schluss Daten statistisch ausgewertet.

Doch egal auf welche Methoden man für eine sportwissenschaftliche Studie zurückgreift – am Ende müssen die Ergebnisse jedoch immer denselben wissenschaftlichen Gütekriterien standhalten, nämlich *Objektivität, Reliabilität, interne Validität, Inhaltsvalidität, externe Validität* und *Replizierbarkeit*. Werden diese Qualitätsstandards nicht ausreichend berücksichtigt, ist eine Studie in ihrer Aussagekraft eingeschränkt – im Extremfall sogar wertlos. In diesem Whitepaper werden daher die wichtigsten Qualitätsstandards einer sportwissenschaftlichen Untersuchung beschrieben und mit Beispielen aus der Praxis illustriert.

INHALT

1. ÜBERBLICK WISSENSCHAFTLICHER GÜTEKRITERIEN

Trainingsmethoden basieren oftmals selbst im Leistungssport auf der jahrelangen Sport- oder Spielpraxis eines Trainers. Anders gesagt: Man trainiert auf eine bestimmte Weise, weil man es schon immer so gemacht hat. Für die Effektivität dieser Trainingsmethoden gibt es jedoch oft keinen empirischen Beleg. Das gilt jedoch auch umgekehrt für so manches neue Trainingsverfahren. Nur weil eine Methode gerade in Mode ist, bedeutet das noch lange nicht, dass sie herkömmlichen Verfahren überlegen ist. Gerade im Profisport, wo das Leistungsniveau insgesamt so hoch ist, dass man sich durch die Optimierung von Feinheiten und Details den Schlüssel für den entscheidenden Leistungszuwachs und Wettkampferfolg verspricht, baut man daher auf die Trainingswissenschaft als eine Teildisziplin der Sportwissenschaft.

Die Trainingswissenschaft befasst sich mit dem Gegenstand des sportlichen Trainings, der Leistungsfähigkeit und des Wettkampfs. Ihr Ziel ist es, Trainingsmethoden zu finden und zu optimieren, mit denen die sportliche Leistung gesteigert wird. Die Trainingswissenschaft ist interdisziplinär angelegt und benutzt Erkenntnisse und Methoden mehrerer anderer wissenschaftlicher Disziplinen wie der Naturwissenschaften, Sozialwissenschaften und der Sportmedizin. Beispielsweise bedient sie sich zur Erfolgskontrolle der Leistungsdiagnostik und bewegt sich auf einem Gebiet, auf dem die Psychologie mit einer ihrer Teildisziplinen, der Differentiellen Psychologie, in mittlerweile über hundert Jahren ein äußert umfangreiches und mächtiges Arsenal an Testmethoden entwickelt hat
(– dass es in der Psychologie um die Messung geistiger Leistungen wie Intelligenz oder Konzentration und im Sport um körperliche Leistungen geht, spielt aus testtheoretischer Sicht keine Rolle).

Um die Wirksamkeit einer bestimmten Trainingsmethode nachzuweisen, verwendet die Trainingswissenschaft Methoden der Sozialwissenschaften: Sie führt beispielsweise ein Experiment durch mit einer Leistungsdiagnostik zu zwei Messzeitpunkten (vorher versus nachher) und der Trainingsmaßnahme zwischen diesen Messungen. Oder sie vergleicht zwei Gruppen von Sportlern, die mit unterschiedlichen Methoden trainieren (Experimentalgruppe versus Kontrollgruppe). Sie operationalisiert Leistung durch geeignete Parameter (z. B. benötigte Zeiten, erzielte Weiten oder Höhen in Metern, Häufigkeiten wie Treffer, Tore oder Fehler) und kommt dadurch zu Daten, die mit statistischen Verfahren analysiert werden.

Doch egal für welche Methode zur Evaluation von Trainingsmethoden sich ein Sportwissenschaftler entscheidet, egal für welches Untersuchungsdesign und für welche Auswertungsmethode – am Ende müssen die Ergebnisse jedoch immer denselben wissenschaftlichen Gütekriterien standhalten, nämlich

- Objektivität
- Reliabilität
- Interne Validität
- Inhaltsvalidität
- Externe Validität
- Replizierbarkeit

Werden diese Qualitätsstandards nicht ausreichend berücksichtigt, ist eine Studie in ihrer Aussagekraft eingeschränkt – im schlimmsten Fall sogar völlig wertlos. In diesem Whitepaper werden daher die wichtigsten Qualitätsstandards einer sportwissenschaftlichen Untersuchung beschrieben und mit Beispielen aus der Praxis illustriert.

2. OBJEKTIVITÄT

Objektivität ist bei einer Messmethode dann gegeben, wenn die mit dieser Methode erzielten Ergebnisse nicht davon abhängen, *wer* die Methode anwendet (Klein et al., 2012). Gibt es beispielsweise zwei verschiedene Testleiter, dann muss sichergestellt werden, dass beide sich gegenüber den Testpersonen gleich verhalten und z. B. identische Instruktionen geben. Ansonsten könnte es sein, dass durch die Art, wie ein Versuchsleiter die Aufgabenstellung erklärt, manche Probanden indirekt nützliche Tipps bekommen und andere nicht, oder dass Einweisungen in die Aufgabe unterschiedlich präzise und verständlich sind. Um eine möglichst hohe *Durchführungsobjektivität* zu gewährleisten, ist es daher empfehlenswert, dass derselbe Testleiter alle Testpersonen instruiert. Manchmal kann es auch hilfreich sein, Instruktionen schriftlich zu geben. In jedem Fall sinnvoll ist eine Schulung und Sensibilisierung der Test- oder Versuchsleiter.

Auch bei der Anwendung von Messtechnik kann es zu Problemen mit der *Durchführungsobjektivität* kommen. Bei einer Pulsuhr mit einem EDA-Sensor (= GSR-Sensor) zur Messung der elektrischen Hautleitfähigkeit (= **e**lektro**d**ermale **A**ktivität), die als Maß für Schwitzen, aber auch für Stress und Erregtheit verwendet wird, hängt das Messergebnis auch vom Druck des Sensors auf der Haut ab. Legen verschiedene Testleiter das Armband einer Person unterschiedlich fest an, führt dies zu abweichenden Messwerten.

Bei der Analyse der erbrachten Leistung oder der aufgenommen Daten muss zudem *Auswertungsobjektivität* gegeben sein. Kritisch ist Auswertungsobjektivität immer dann, wenn Leistung nicht auf simpel auswertbare Variablen wie Längenmaße, Zeiten oder Häufigkeiten reduzierbar ist, sondern beispielsweise komplexe Bewegungsabläufe beurteilt werden müssen wie beim Kunstturnen, Eiskunstlauf oder Boxen, wo Punktrichter Noten vergeben. Auch Verhaltensanalysen wie zum Beispiel von der Spielweise einer Fußballmannschaft müssen unabhängig davon sein, welcher Beurteiler das Spielverhalten analysiert. Ab wann ein Ballverlust einer Mannschaft als Fehlpass zu werten ist oder ein Ball Richtung gegnerisches Tor als Torschuss, ist in der Praxis komplizierter als es auf den ersten Blick scheint. Entsprechend präzise müssen daher auch die Regeln zur Codierung des beobachteten Verhaltens sein, um *Objektivität* sicherzustellen.

Abbildung 1: Beim Eiskunstlauf wird Leistung gemessen, indem Punktrichter Noten vergeben. Durch statistische Verfahren lässt sich überprüfen, wie einig sich dabei die Urteiler sind und inwiefern dabei Auswertungsobjektivität gegeben ist.

Wie objektiv ein Messverfahren ist, lässt sich in vielen Fällen auch überprüfen und in Zahlen ausdrücken. Hierfür gibt es verschiedene statistische Maße. Beispiel Eiskunstlauf: Das *ISU Wertungssystem für Eiskunstlauf und Eistanzen* verwendet eine 10stufige Wertungsskala für jeden Programmbestandteil einer Darbietung. Mehrere Preisrichter beurteilen dabei die Darbietungen aller Teilnehmer. Für jeden Programmbestandteil resultiert somit eine Datenmatrix, die so viele Spalten wie Preisrichter und so viele Zeilen wie Teilnehmer hat. Naturgemäß sind sich die Preisrichter bei ihren Punktevergaben keineswegs immer einig, sondern ihre Urteile differieren. Über eine sogenannte Reliabilitätsanalyse ist es über die Berechnung eines *Cronbachs Alpha* möglich, die Einigkeit unter den Preisrichtern in einer einzigen Zahl zusammenzufassen: Ein Wert von 0 bedeutet, dass sich die Urteiler überhaupt nicht einig sind und ihre Urteile per Zufall vergeben, 1 bedeutet, dass sie in ihren Urteilen vollkommen übereinstimmen. Als Faustregel lässt sich sagen: Ein *Cronbachs Alpha* > 0,9 ist exzellent, > 0,8 gut, > 0,7 akzeptabel, > 0,6 fragwürdig und alles darunter schlecht (George & Mallery, 2002). Dieselbe Reliabilitätsanalyse ermöglicht es auch über die Berechnung von Trennschärfekoeffizienten, mit einem Blick zu sehen, welcher Preisrichter mit dem Gesamturteil *aller* Preisrichter wie gut übereinstimmt. Berechnet wird dies jeweils über die Korrelation der Urteile eines bestimmtes Preisrichters mit dem Gesamturteil *aller* Preisrichter (= Gruppenstandard). So lässt sich leicht ein Urteiler herausfinden, der einen völlig anderen Geschmack oder andere Auffassungen

von einer gelungenen Performance hat als die Mehrheit der Preisrichter. Man erkennt ihn daran, dass seine Urteile nur schwach oder sogar negativ mit dem Gesamturteil korrelieren.

Bei anderen Urteilen im Sport geht es nicht um eine *graduelle* Abstufung auf einer Skala, sondern um *kategoriale* Urteile, z. B. Foul oder Nicht-Foul, Abseits oder Nicht-Abseits. Auch hier können verschiedene Urteiler zu unterschiedlichen Resultaten kommen. Und auch hier kann man das Ausmaß der Urteilsübereinstimmung zwischen zwei Urteilern leicht berechnen (= *Interrater-Reliabilität*). Als statistisches Maß bei kategorialen Daten verwendet man hierfür das sogenannte *Cohen's Kappa* (Cohen, 1960; Landis & Koch, 1977; Bortz, Lienert & Boehnke, 1990), das die Übereinstimmung von zwei Urteilern als Zahl zwischen 0 und 1 ausgibt. Zur Berechnung der Übereinstimmung von mehr als zwei Urteilern gibt es das sogenannte *Fleiss' Kappa* (Fleiss, 1971; Fleiss, 1981). Die Interpretation der Kappa-Werte ist ähnlich wie bei den *Cronbachs Alpha*-Werten (siehe oben!).

Obwohl solche Analysen mit einer Statistik-Software kaum Aufwand bedeuten und mit wenigen Mausklicks erledigt sind, werden sie in der Praxis der Sportwissenschaften kaum durchgeführt, möglicherweise weil es Sportwissenschaftlern oft am methodischen Problembewusstsein fehlt. Dies ist nicht weiter erstaunlich, da im Sport Eigenschaften und Verhaltensweisen gemessen werden, die verglichen mit anderen Disziplinen (z. B. der Psychologie) relativ leicht objektiv und präzise messbar sind. Daher erscheint es Sportwissenschaftlern als wenig lohnend, sich mit dem komplizierten Thema der Messmethodik auseinanderzusetzen.

Abbildung 2: Die Verwendung von technischen Messgeräten garantiert noch keine Objektivität, solange diese von Menschen bedient werden.

Doch selbst bei solchen Disziplinen, in denen einfache Größen wie Längenmaße oder Zeiten gemessen werden, bedeutet das nicht, dass die Messergebnisse automatisch *objektiv* sind. Beispiel: Bei einem Hundertmeter-Lauf werden zwar Zeiten mit einer Stoppuhr gemessen, doch dieses Messinstrument wird von einem Menschen bedient. Je nachdem, wie er die Stoppuhr bei Start und Ende drückt, weicht die gemessene Zeit von der tatsächlich benötigten Zeit des Läufers ab. Beweisen lässt sich dies, indem man zwei Beobachter mit jeweils einer Stoppuhr ausstattet und sie unabhängig voneinander die Läufe von beispielsweise 50 Läufern stoppen lässt. Bei jedem Lauf werden sich die beiden Messungen geringfügig unterscheiden. Bildet man für jeden Lauf die Differenz dieser beiden Messwerte, erkennt man jeweils den Einfluss der Messung. Der Mittelwert aus den

Beträgen dieser Differenzen fasst diese Ungenauigkeit zu einer einzigen Zahl zusammen. Wenn sich zeigt, dass der Einfluss des Messenden so groß ist, dass er die zu messenden Variablen in erheblichem Maß verfälscht, sollten Maßnahmen getroffen werden, um diesen Einfluss zu reduzieren. Beispielsweise können Testleiter geschult werden, oder es kann möglicherweise der Faktor Mensch sogar komplett aus dem Messvorgang eliminiert werden. Wo zum Beispiel eine Lichtschranke Zeiten messen kann, müssen auch keine Menschen mehr Stoppuhren bedienen, und das *Objektivitätsproblem* ist gelöst.

3. RELIABILITÄT

Die *Reliabilität* (= Zuverlässigkeit) gibt den Grad der Messgenauigkeit (= Präzision) einer Messmethode an (Bortz & Döring, 2006). Jede Messung ist auch immer mit einem Messfehler behaftet. Je kleiner dieser Messfehler, desto höher ist die *Reliabilität*. Das Angenehme an der *Reliabilität* ist, dass man sie – anders als *interne* oder *externe Validität* – quantifizieren kann. Denn wenn man beispielsweise zweimal dasselbe misst, muss auch zweimal dasselbe Ergebnis rauskommen. Das ist jedoch in der Praxis keineswegs immer der Fall.

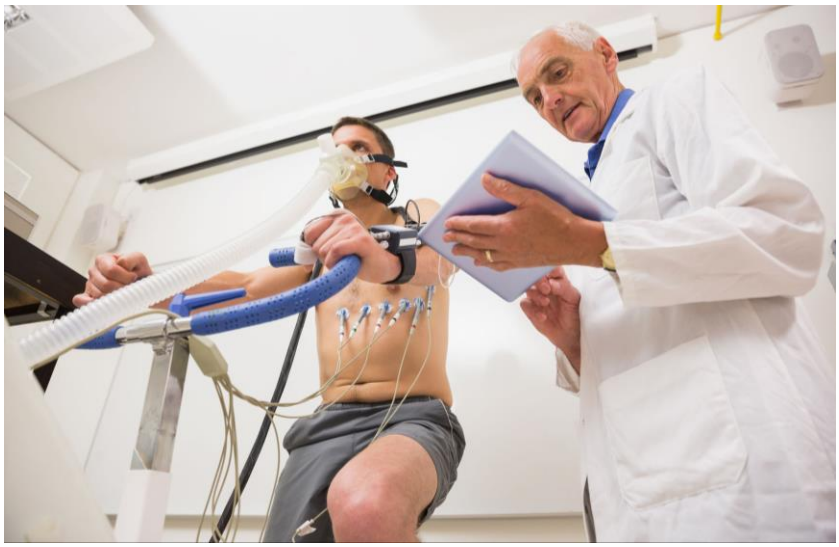


Abbildung 3: Pulsuhren messen Körperfunktionen wie die Herzrate nur mit begrenzter Genauigkeit erfassen. Zuverlässiger ist ein EKG.

Verwendet man beispielsweise eine Pulsuhr, um basierend auf Herzrate oder Herzratenvariabilität Rückschlüsse auf die Fitness eines Sportlers zu ziehen, dann steckt in diesem gemessenen Wert immer auch ein mehr oder weniger großer Messfehler. Das beginnt bereits damit, dass ein *Photoplethysmographie-Sensor* (*PPG*, das rote Licht, das die Haut durchleuchtet) am Handgelenk die Herzrate nicht so exakt messen kann wie beispielsweise auf den Brustkorb geklebte EKG-Sensoren bei einem Medizin-Gerät. Beweisen lässt sich dies über einen sogenannten *Paralleltest*, indem ein Sportler auf einem Laufband am Handgelenk eine Pulsuhr trägt und gleichzeitig mit einem medizinisch zugelassen Gerät EKG-Daten aufgezeichnet werden. Das Medizingerät dient dabei als Referenz (Mell, 2010). Die Korrelation der von einer Pulsuhr aufgezeichneten Datenreihe

(über die Zeit) mit der vom Medizingerät aufgezeichneten Datenreihe zur Herzrate gilt dabei als Maß für die *Reliabilität* der Pulsuhr. Im Idealfall ist die Korrelation 1, ein Wert von 0,7 sollte stets mindestens erreicht werden.

Auch Pulsuhren verschiedener Hersteller, die mit demselben PPG-Sensor arbeiten, messen keineswegs genau gleich, weil sich die dahinterliegenden Auswertealgorithmen der Pulsuhren unterscheiden. Beweisen lässt sich dies, indem derselbe Sportler an den Handgelenken zwei verschiedene Pulsuhren trägt. Auch hier korrelieren die Datenreihen nie mit 1, sondern abhängig von den verwendeten Pulsuhren deutlich niedriger.

Abbildung 4: Bei jedem Wurf eines Kugelstoßers spielt nicht nur seine Leistungsfähigkeit eine Rolle für die erzielte Weite, sondern auch der Zufall. Dies senkt die Reliabilität einer Messung, lässt sich jedoch mit den richtigen Messverfahren in den Griff bekommen.

Doch nicht nur Technik ist in ihrer Genauigkeit begrenzt. Bereits bei der Durchführung des Messverfahrens hat man es mit Messfehlern zu tun. Besonders deutlich wird dies bei einem Sporteignungstest. Bei solch einem Eignungstest (z. B. für die Zulassung zu einem Sportstudium) soll ein Urteil über die Leistungsfähigkeit einer Person hinsichtlich verschiedener Disziplinen getroffen werden. Das Problem dabei: Beobachtbar und damit messbar sind immer nur *Verhaltensweisen* einer Person, nicht ihre Leistungsfähigkeit selbst. Beim Kugelstoßen beispielsweise stößt ein Sportler die Kugel nicht bei jedem Versuch gleich weit. Denn bei jeder Ausführung eines Stoßes spielt auch der Zufall eine Rolle. Von der Beobachtung eines einzigen Stoßes auf die Leistungsfähigkeit des Kugelstoßers zu schließen, wäre daher sehr ungenau, also *wenig reliabel*. Beweisen lässt sich dies, indem man zum Beispiel 50 Sportler eine Kugel in zwei Durchgängen stoßen lässt und die erzielten Weiten misst *(= Retest-Reliabilität)*. Man erhält eine Datenmatrix mit 50 Zeilen und 2 Spalten. Anschließend berechnet man die Korrelation dieser beiden Spalten. Die Korrelation ist dabei keineswegs 1, sondern vielleicht 0,5 bis 0,8 – umso geringer, je ähnlicher das Leistungsniveau der Sportler und je größer der Zufallsfaktor ist. Von denjenigen, die beim ersten Versuch zu den 10 besten gehörten, sind vielleicht beim zweiten Versuch nur noch 4 oder 5 darunter. Diesen Effekt bezeichnet man als *Regression zum Mittelwert* – er ist umso größer, je größer der Zufallseinfluss bei einer Messung ist (Bortz & Döring, 2006).

Möchte man die Leistungsfähigkeit beim Kugelstoßen möglichst zuverlässig diagnostizieren, ist ein Ausweg, die Anzahl der Messungen zu erhöhen. Jeder Sportler bekommt beispielsweise 10 Versuche. Denn redundante Messungen erhöhen die *Reliabilität*. Gewertet wird dann entweder der beste Versuch eines Kandidaten (= Maximalwert) oder der Durchschnittswert. Eine solchermaßen ermittelte Weite ermöglicht einen zuverlässigeren Rückschluss auf die Leistungsfähigkeit eines Sportlers in dieser Disziplin als die Auswertung einer einzelnen Wurfweite.

4. INTERNE VALIDITÄT

Validität bedeutet allgemein Gültigkeit. Sie gibt an, ob ein Messverfahren das misst, was es messen soll. Die *interne Validität* betrifft die Frage, wie gut es in einer Untersuchung gelingt, bedeutsame Störvariablen zu kontrollieren. Eine Untersuchung ist also dann *intern valide*, wenn ihre Ergebnisse eindeutig interpretierbar sind (Campbell & Stanley, 1963; Cook & Campbell, 1979).

Möchte man beispielsweise eine bestimmte motorische Fähigkeit wie die Schnelligkeit eines Sportlers messen, ergibt sich das Problem, dass Schnelligkeit oft von anderen motorischen Fähigkeiten (Kraft, Ausdauer, Gelenkigkeit oder Koordination) beeinflusst ist. Ziel muss daher sein, das Testverfahren so zu gestalten, dass diese anderen Einflussfaktoren möglichst ausgeschaltet sind. Andersfalls könnte es sein, dass eine geringe gemessene Laufgeschwindigkeit nicht an Defiziten bei der Schnelligkeit, sondern in Wirklichkeit an mangelnder Ausdauer liegt. Man spricht dann von einer Konfundierung dieser beiden Variablen. Das Ergebnis wäre dann nicht mehr eindeutig interpretierbar und die Messung dadurch *intern nicht valide*.

Abbildung 5: Schnelligkeit, Ausdauer, Gelenkigkeit, Koordination – beim Hürdenlauf kommt es auf verschiedene Dinge an. Für die valide Messung einzelner motorischer Fähigkeiten müssen ihre Wechselwirkungen entsprechend berücksichtigt werden.

Auch andere Störfaktoren können die Messung motorischer Fähigkeiten verfälschen, beispielsweise Erwärmung der Muskeln, emotionale Erregung, körperliche Ermüdung oder Umgebungsbedingungen wie Wind- oder Temperaturverhältnisse. Daher gibt es für die Planung einer Serie von motorischen Tests die Faustregel für die inhaltliche Reihenfolge: (1) Technik stets vor anderen Inhalten, (2) schwierige technische Inhalte vor einfachen Inhalten und (3) Technik vor Schnelligkeit vor Kraft vor Ausdauer.

5. INHALTSVALIDITÄT

Inhaltsvalidität (auch *Face Validity*, Augenscheinvalidität oder Logische Validität) ist gegeben, wenn das, was man eigentlich messen möchte, von der Messmethode in seinen wichtigsten Aspekten erschöpfend erfasst wird (Bortz & Döring, 2006; Rosnow & Rosenthal, 2007).

Um beispielsweise für eine Ausbildung zum Polizisten zugelassen zu werden, ist es Voraussetzung, einen Sporteignungstest zu bestehen. Hintergrund ist, dass ein Polizeibeamter körperlich so fit sein soll, dass er auch einen flüchtigen Straftäter stellen kann. Doch darüber, welche konkreten körperlichen Fähigkeiten dafür wichtig sind und Bestandteil eines Eignungstests sein sollten, kann man durchaus unterschiedlicher Meinung sein. Gehört Schwimmen dazu? Oder sogenannte „Kleinbanksprünge auf Zeit“ (http://www.sporttest-polizei.de/)?

Abbildung 6: „Kleinbanksprünge auf Zeit“ – ein Bestandteil des Sporttests für die Zulassung zur Polizeiausbildung. Bildquelle: Main-Post

Auch zur Messung der Leistung einer Fußballmannschaft werden heutzutage allerlei statistische Größen gemessen (zum Beispiel zurückgelegte Strecke pro Spieler, Anzahl der Torschüsse, Eckstöße, Ballbesitz oder der Anteil gewonnener Zweikämpfe). Auch über die Aussagekraft dieser Maße kann man unterschiedlicher Meinung sein. So macht es beispielsweise einen Unterschied, ob der Ballbesitz überwiegend in der eigenen Spielfeldhälfte oder der des Gegners stattfindet. Letztlich kann man die *Inhaltsvalidität* einer Messung nicht – wie beispielsweise die Reliabilität – mit statistischen Kennzahlen belegen, sondern nur inhaltlich, argumentativ und theoriegeleitet. Ein Urteil über Inhaltsvalidität bleibt daher Experten für den jeweiligen Bereich vorbehalten. Und da auch Experten häufig verschiedene Ansichten vertreten, bleibt dabei auch immer Raum für Meinungsverschiedenheiten und Diskussionen.

BUNDESLIGA - STATISTIK

Torjäger | Scorer | Karten | Torschüsse | Zweikämpfe | Ballkontakte | Pässe | Gefoult | Fouls verübt

Platz	Name	Gespielte Minuten	pro Minute	Gesamt
1	Xabi Alonso	1901	1,36	2584
2	David Alaba	2257	1,14	2574
3	İlkay Gündoğan	1993	1,18	2344
4	Mats Hummels	2251	1,03	2310
5	Granit Xhaka	2015	1,14	2300
6	Julian Weigl	2012	1,13	2279
7	Philipp Lahm	2020	1,08	2181
8	Arturo Vidal	1868	1,16	2167
9	Pascal Groß	2473	0,87	2147
10	Jonas Hector	2589	0,81	2100
11	Lewis Holtby	2554	0,8	2033
12	Wendell	2317	0,85	1981
13	Rafinha	1539	1,26	1942
14	Emiliano Insúa	2692	0,72	1930
15	Andreas Christensen	2430	0,79	1922
	Niklas Süle	2700	0,71	1922
17	Vladimír Darida	2426	0,79	1912

Abbildung 7: Ballkontakte-Ranking als Beispiel für ausufernde Statistiken im Fußball. Nicht alles, was sich zahlenmäßig erfassen lässt, sagt deswegen auch etwas über Leistungsfähigkeit oder gar sportlichen Erfolg aus.

6. EXTERNE VALIDITÄT

Eine Untersuchung ist *extern valide*, wenn ihre Ergebnisse über die besonderen Bedingungen der Untersuchungssituation hinaus generalisierbar sind. Man unterscheidet hierbei zwei Formen der Generalisierbarkeit, nämlich eine Generalisierung über *Personen* und *Situationen* (Hager & Westermann, 1983; Cook & Shadish, 1994).

Sollen die Ergebnisse über die konkret in der Studie untersuchten *Personen* hinaus verallgemeinert werden, muss die untersuchte Probandenstichprobe repräsentativ sein. Hat man beispielsweise für eine Untersuchung zur Effektivität einer bestimmten Trainingsmethode wegen der leichteren Verfügbarkeit nur Studierende untersucht, lassen sich die Ergebnisse kaum auf andere Personengruppen übertragen, z. B. auf Leistungssportler, da diese über ganz andere körperliche Voraussetzungen verfügen. Generell gilt: Je geringer die Repräsentativität der untersuchten Probanden-Stichprobe, desto geringer ist die *externe Validität*.

Sollen Ergebnisse über die konkret untersuchte *Situation* hinaus verallgemeinerbar sein, setzt dies beispielsweise voraus, dass sich Sportler in einer Trainingssituation, in der der Erfolg einer Trainingsmethode gemessen wird, genauso verhalten (also beispielsweise dieselben Bewegungsmuster abrufen) wie in einer Wettkampfsituation. Wenn jedoch in der Wettkampfsituation andere Faktoren als in einer standardisierten Trainingssituation das Verhalten beeinflussen (z. B. durch Nervosität oder Konzentration auf gegnerische Mitspieler), dann ist die Testsituation der Studie wenig aussagekräftig für den Wettkampf. Generell gilt: Je unnatürlicher und praxisferner die Untersuchungsbedingungen, desto geringer ist die *externe Validität*.

Ein wichtiger Begriff ist auch die *Kriteriumsvalidität*: Sie bezieht sich auf den Zusammenhang zwischen den Ergebnissen eines Testverfahrens und einem als entscheidend erachteten empirischen Kriterium. Beispielsweise können sich bestimmte Trainingsmethoden auf konkrete Bewegungsabläufe oder physiologische Parameter auswirken, die mit den entsprechenden Methoden auch gut (d. h. *intern valide* und *reliabel*) gemessen werden können. Doch das eigentliche Ziel im Sport ist oft ein anderes, beispielsweise der Erfolg. Setzt zum Beispiel ein Fußballtrainer darauf, durch intensiveres Konditionstraining die Fitness seiner Mannschaft zu verbessern, geht es ihm ja eigentlich nicht darum, die zurückgelegten Kilometer pro Spieler zu erhöhen, sondern Spiele zu gewinnen. Als Kriterium für die Wirksamkeit des Konditionstrainings kann man daher auch die Anzahl erreichter Punkte oder den Tabellenplatz am Saisonende heranziehen. Ist dieser am Saisonende *nach* Einführung des Trainings höher als in der vorherigen Saison *vor* Einführung des Trainings? Erreichen Mannschaften, die mit dieser Methode trainieren, im Durchschnitt eine höhere Punktzahl als solche, die es nicht tun?

Doch so plausibel es auch erscheint, den sportlichen Erfolg als entscheidendes Kriterium heranzuziehen – es sind auch Nachteile damit verbunden: Denn der Erfolg in einem Wettkampf hängt von sehr vielen Faktoren ab (= Multikausalität). Und diese Faktoren sind nicht konstant, d. h. Fußballmannschaften unterscheiden sich voneinander noch in vielen weiteren Eigenschaften und nicht nur beim Konditionstraining. Und selbst bei derselben Mannschaft unterscheiden sich zwei Saisons in vielen Aspekten. Aus methodischer Sicht sind alle diese Faktoren Störvariablen, wenn es darum geht, den Einfluss einer bestimmten Variable wie Konditionstraining zu analysieren. Darum ist es sehr wahrscheinlich, dass sich ein positiver Effekt einer bestimmten Trainingsmethode beim Fußball auf den sportlichen Erfolg wissenschaftlich *nicht* nachweisen lässt. Ein möglicher kleiner positiver Effekt geht einfach im Rauschen der zahlreichen Störvariablen unter.

Aus diesem Grund greift man dann eben doch für die Evaluierung der Effektivität von Trainingsmethoden auf abstraktere Kriterien zurück wie zum Beispiel beim Konditionstraining die zurückgelegte Strecke eines Spielers im Spiel oder physiologische Leistungsparameter. Dies erhöht somit die *interne Validität*, senkt aber die *externe Validität*, da Fußball eben kein Langstreckenlauf ist. Lässt sich tatsächlich damit nachweisen, dass dieses Konditionstraining zu besseren Laufleistungen führt, bleibt dadurch dennoch offen, ob eine Mannschaft deswegen auch häufiger gewinnt. Es ist ein typischer Zielkonflikt: *Externe Validität* geht meist zulasten von *interner Validität* – und umgekehrt. Beide Qualitätskriterien gleichzeitig zu optimieren, muss zwar das Ziel sein. Es bleibt jedoch ein Ideal, so dass man sich in der Regel immer mit einem Kompromiss zufriedengeben muss (Bortz & Döring, 2006).

Wenn die Beobachtung oder Messung das zu beobachtende oder zu messende Verhalten und somit auch die erfassten Daten beeinflussen, spricht man auch von *Reaktivität*. Diese mindert die *externe Validität*. *Reaktivität* ist damit ein Negativ-Kriterium. Bei Beobachtungsstudien ist dieses Phänomen seit fast 100 Jahren unter dem Begriff *Hawthorne-Effekt* bekannt (Roethlisberger & Dickson, 1964). Dieser besagt, dass die Teilnehmer einer Studie ihr natürliches Verhalten ändern, weil sie wissen, dass sie an einer Studie teilnehmen und unter Beobachtung stehen. Im Sportbereich bedeutet dies, dass sich Studienteilnehmer beispielsweise auf die auszuführenden Bewegungsabläufe konzentrieren und so ausführen, wie vom Trainer gewünscht. Das Wesen der Motorik ist jedoch, dass sie hoch automatisiert ist, d. h. dass Bewegungsabläufe eben *keiner* besonderen Aufmerksamkeit mehr bedürfen, sobald sie einmal erlernt und verinnerlicht sind.

Reaktivität im Bereich Sportwissenschaften äußert sich daher manchmal darin, dass Studienteilnehmer im Training neu erlernte, aber noch nicht vollständig automatisierte, Bewegungsabläufe durch erhöhte kognitive Kontrolle genau dann zeigen, wenn sie wissen,

dass sie unter Beobachtung stehen, jedoch danach wieder in ihre gewohnten Bewegungsmuster zurückfallen. Auch kann es sein, dass sie sich insgesamt mehr anstrengen, sobald sie wissen, dass ihre Leistung gemessen wird. Dadurch können sie insgesamt ebenfalls eine höhere Performance erzielen, die mit einem bestimmten Bewegungsmuster (d. h. mit der unabhängigen Variablen der Studie) noch nicht einmal etwas zu tun haben muss.

Begegnen kann man dieser Tendenz zur *Reaktivität* zum Beispiel, indem man Verhaltensbeobachtungen oder Messungen verdeckt ohne das Wissen der Sportler durchführt – sofern dies möglich ist – oder indem man die Dauer der Verhaltensbeobachtung stark ausweitet. Denn sich über einen langen Zeitraum zu „verstellen", ist deutlich schwieriger als für kurze Zeit. Durch die somit geringere *Reaktivität* steigt die *externe Validität* einer Studie.

7. REPLIZIERBARKEIT

Eine Replikation ist eine Wiederholung einer Untersuchung unter identischen oder sehr ähnlichen Bedingungen. Wiederholt man beispielsweise eine Studie zum leistungssteigernden Effekt einer bestimmten Trainingsmethode mit einer neuen Stichprobe an Teilnehmern, die die gleichen körperlichen Voraussetzungen mitbringen wie die Teilnehmer der ersten Studie, und kommen dabei dieselben oder sehr ähnliche Ergebnisse heraus wie beim ersten Mal, dann sind die Schlussfolgerungen aus der Studie deutlich besser abgesichert als ohne die Wiederholung. Gelingt es jedoch trotz gleicher Voraussetzungen *nicht*, die Ergebnisse zu reproduzieren, dann ist die Studie *nicht replizierbar*. Dies ist ein Hinweis darauf, dass in der Durchführung oder Auswertung der Studien – also auch schon beim ersten Mal – irgendwelche Störvariablen das Ergebnis verfälscht haben, z. B. Mängel bei der *Objektivität*, *nicht reliable* Messungen, unkontrollierte Störvariablen (also mangelnde *interne Validität*) oder auch Fehler bei der statistischen Datenanalyse.

Der Grundsatz, dass wissenschaftliche Untersuchungsergebnisse reproduzierbar sein müssen, ist ein fundamentales Prinzip der Wissenschaft. Können Ergebnisse von anderen unabhängigen Forschergruppen nicht reproduziert werden, oder kommen diese sogar zu gegenteiligen Ergebnissen, bedeutet es auf lange Sicht das Aus für die Schlussfolgerungen aus solchen Studien.

Um die *Replizierbarkeit* sozialwissenschaftlicher Studien ist es generell nicht sonderlich gut bestellt, und gerade in den letzten Jahren ist eine hitzige Debatte unter Wissenschaftlern entstanden, woran das liegt und was sich dagegen tun lässt (Francis, 2012; Pashler & Harris, 2012; Giner-Sorolla, 2012; Ioannidis, 2012). Es gibt keinen Grund zur der Annahme, dass sportwissenschaftliche Untersuchungen von diesem Problem nicht betroffen seien.

Echte Replikationsstudien gibt es in den Sportwissenschaften sehr selten. Identisch sind in der Regel höchstens Fragestellungen oder auch Trainingsmethoden, aber nicht der gesamte Untersuchungskontext, der einen großen Einfluss auf das Ergebnis hat (Makel, Plucker & Hegarty, 2012). Zudem fehlt sowohl der Anreiz, die Studien anderer Forscher identisch zu wiederholen, als auch eigene Ergebnisse kritisch zu hinterfragen (Koole & Lakens, 2012).

8. FAZIT

Die beschriebenen Kriterien sind die Qualitätsstandards für jede Art von wissenschaftlichen Untersuchungen und gelten nicht nur für sportwissenschaftliche Untersuchungen. Alle Kriterien sind wichtig. Daher darf es nicht passieren, dass ein oder mehrere Kriterien vernachlässigt werden. In der Praxis ist dies jedoch nicht leicht zu realisieren, weil die Optimierung eines Kriteriums häufig zu einer Verschlechterung bei einem anderen Kriterium führt. Beispielsweise führt die Maximierung der *internen* Validität häufig zu einer Reduzierung der *externen* Validität und umgekehrt. Das sollte jedoch nicht entmutigen, sondern im Gegenteil ein umso größerer Ansporn sein, das Bestmögliche aus den gegebenen Rahmenbedingungen herauszuholen.

Denn all diese Qualitätskriterien sind nicht etwa Forderungen weltfremder Wissenschaftler mit einem Hang zu Idealismus und methodischer Besserwisserei. Sie dienen vielmehr alle demselben Zweck, nämlich die Wahrheit herauszufinden, also beispielsweise die *tatsächlichen* Ursache-Wirkungsbeziehungen zu bestimmten Trainingsmethoden und sportlicher Leistung zu erkennen. Gerade in einem so angewandten und praxisorientierten Bereich wie den Sportwissenschaften ist daher der praktische Nutzen einer sauberen Methodik evident: Die so gewonnen Studienergebnisse sind dadurch auch wirklich aussagekräftig und belastbar. Bei neuen Trainingsmethoden lässt sich somit evidenzbasiert die Spreu vom Weizen trennen und gewährleisten, dass neu eingesetzte Trainingsverfahren auch tatsächlich effektiver sind als die gewohnten alten. Gerade im Leistungssport, wo es um die Optimierung von Details geht, kommt damit der Sportwissenschaft eine Schlüsselrolle zu, die für den sportlichen Erfolg den Ausschlag geben kann.

9. LITERATUR

Bortz, J. & Döring, N. (2006). Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler. Heidelberg: Springer.

Bortz, J., Lienert, G. A. & Boehnke, K. (1990). Verteilungsfreie Methoden in der Biostatistik. Berlin: Springer.

Campbell, N. & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37-46.

Cook, T. D. & Campbell, D. T. (1979). Quasi-experimentation. Boston: Houghton Mifflin.

Cook, T. D. & Shadish, W. R. (1994). Social experiments: Some developments over the past fifteen years. Annual Review of Psychology, 45, 545-580.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76,* 378-382.

Fleiss, J. L. (1981). The measurement of interrater agreement. In J. L. Fleiss, B. Levin & M. C. Paik., *Statistical methods for rates and proportions* (pp. 212-236), New York: John Wiley & Sons.

Francis, G. (2012). The psychology of replication and replication in psychology. Perspectives on Psychological Science, 7, 585-594.

George, D & Mallery, P. (2002). *SPSS for Windows Step by Step: A Simple Guide and Reference 15.0 Update.* Pearson.

Giner-Sorolla, R. (2012). Science or art? How aesthetic standards grease the way through the publication bottleneck but undermine science. Perspectives on Psychological Science, 7, 562-571.

Hager, W. & Westermann, R. (1983). Planung und Auswertung von Experimente. In J. Bredenkamp & H. Feger (Hrsg.), Hypothesenprüfung. Göttingen: Hogrefe.

Ioannidis, J. P. A. (2012). Why science is not necessarily self-correcting. Perspectives on Psychological Science, 7, 645-654.

Klein, O., Doyen, S., Leys, Ch., Magalhães de Saldanha da Gama, P. A., Miller, S., Questienne, L. & Cleeremans, A. (2012). Low hopes, high expectations: Expectancy effects and the replicability of behavioral experiments. *Perspectives on Psychological Science, 7,* 572-584.

Koole, S. & Lakens, D. (2012). Rewarding replications: A sure and simple way to improve psychological science. Perspectives on Psychological Science, 7, 608-614.

Landis, J. R. & Koch G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33, 159-174.

Makel, M. C., Plucker, J. A. & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? Perspectives on Psychological Science, 7, 537-542.

Mell, J. (2010). *Zuverlässigkeit und Genauigkeit von Pulsoximetern der dritten und vierten Generation unter besonderer Berücksichtigung des Alarmierungsverhaltens im klinischen Gebrauch*. Dissertation. Erlangen [pdf-Dokument, verfügbar unter: https://opus4.kobv.de/opus4-fau/frontdoor/index/index/docId/1487

Pashler, H. & Harris, Ch. R. (2012). Is the replicability crisis overblown? Three arguments examined. Perspectives on Psychological Science, 7, 531-536.

Roethlisberger, F. J. & Dickson, W. J. (1964). Management and the worker. Cambridge, Mass.: Harward University Press.

Rosnow, R. L & Rosenthal, R. (2007). Beginning behavioral research. A conceptual primer. Upper Saddle River, NJ: Pearson Education.

Die Ergoneers GmbH wurde 2005 als Spin-off des Lehrstuhls für Ergonomie der Technischen Universität München gegründet. Heute ist das Unternehmen mit weltweiten Standorten in Deutschland (Hauptsitz bei München) und USA sowie zahlreichen Vertriebspartnern ein international wichtiger Partner für die Branchen Transport und Automotive, Marktforschung und Nutzerfreundlichkeit (Usability), Wissenschaft und Forschung sowie Sport und Biomechanik.

Neben der Entwicklung, Herstellung und dem Vertrieb von Mess- und Analysesystemen zur Erforschung von Verhalten und zur Optimierung der Interaktion von Mensch und Maschine bietet Ergoneers umfassende Kompetenz in allen Phasen des Studienablaufs. Zur Ergoneers-Produktpalette zählt vor allem die 360-Grad-Lösung D-LAB, eine umfassende Erfassungs- und Auswertungsplattform für Nutzer- und Verhaltensstudien, mit deren Software-Modulen sich Daten in den Bereichen Eye-Tracking, Datastream, Video, Audio, Physiologie und CAN-Bus messen und analysieren lassen. Mit dem Ergoneers-Blickerfassungssystem Dikablis liefert Ergoneers zudem die passende Hardware, um professionelles Eye-Tracking im realen oder virtuellen Umfeld zu betreiben.

Ergoneers Group
Gewerbering 16
82544 Egling
Germany

T +49.8176.99894-0
F +49.8176.99894-15

Ergoneers of North America, Inc.
111 SW 5th Ave
Suite 3150
Portland, OR 97204
USA

T +1.503.444.3430

info@ergoneers.com
www.ergoneers.com

ERGONEERS
FROM SCIENCE TO INNOVATION