Erwin Tschirner

# Validierung als Argumentation

Die aktuelle Diskussion in der Testforschung fokussiert immer stärker die zentrale Bedeutung der Validität von Testergebnissen, wobei Validität nicht als eine Eigenschaft eines Tests gesehen wird, sondern als das, was mit Testergebnissen gemacht wird, die Konsequenzen, die Tests und Testergebnisse für ganz unterschiedliche Interessensgruppen haben. Auf der Basis von Toulmins (1958) Argumentationsmodell entwickelten Bachman und Palmer (2010) Kane (2006) folgend ein argumentbasiertes Modell von Validierung, in dem Fakten, Behauptungen und Schlussregeln auf mehreren Ebenen des Validierungsprozesses aufeinander bezogen werden müssen. Nach einem sehr kurzen Blick auf die Rolle der Validität in der klassischen Testtheorie, stellt dieser Beitrag zunächst Toulmins Argumentationsmodell vor und beschreibt im Anschluss daran detailliert Bachman und Palmers argumentbasierten Ansatz zur Validierung von Sprachtests, in dem Validierung als die datengestützte Begründung von Behauptungen auf fünf Ebenen gesehen wird, wobei die Schlussfolgerungen einer Ebene als Daten für die nächste Ebene genutzt werden. Dabei ist zu zeigen, dass die Performanz von Testteilnehmern sich korrekt in Testergebnissen widerspiegelt, dass diese Ergebnisse im Hinblick auf die Lebenswirklichkeit der Testteilnehmer interpretiert werden können, dass auf Grundlage dieser Interpretationen faire Entscheidungen getroffen werden können und schließlich, dass die Konsequenzen dieser Entscheidungen positiv für alle Beteiligten sind.

SCHLÜSSELWÖRTER: Testforschung, Validität, Validierung, Impact, Argumentationsmodell, Toulmin.

Erwin Tschirner
Herder-Institut,
Universität Leipzig

Artículo recibido el 24/02/2014 y aceptado el 03/04/2014
VERBUM ET LINGUA
NÚM. 3
ENERO / JUNIO 2014
ISSN 2007-7319

Current language test validation research focuses increasingly on the social consequences of the uses of test scores. In this line of reasoning, validity is no longer considered a property of tests but rather a property of what is done with test scores, the impact test scores and their use have on stake holders including test takers. On the basis of Toulmin's (1958) Model of Argumentation, Bachman and Palmer (2010) following Kane (2006, 2012) developed the argument model

towards validation in which data, claims, and warrants need to be interrelated at various levels of the validation process. After a very brief overview of the role of validity in classical test theory, this paper presents Toulmin's Model of Argumentation and then proceeds to explain Bachman and Palmer's (2010) argument-based approach towards language test validation in which claims are based on data, warrants, and backings at five levels ordered hierarchically, moving from test tasks to test scores to the interpretations of test scores to the decisions that are based on these interpretations, and finally to the consequences these decisions have.

KEY WORDS: Test Research, Validity, Validation, Impact, Argument Model, Toulmin.

## Einleitung

Die klassische Testtheorie kennt drei Gütekriterien: Reliabilität, Objektivität und Validität. Die klassische Testtheorie geht davon aus, dass jede Messung einem Messfehler unterworfen ist und versucht diesen Messfehler so gering wie möglich zu halten. Es gibt einen gemessenen Wert und einen wahren Wert. Messfehler bewirken, dass sich der gemessene Wert vom wahren Wert unterscheidet.

Unter Reliabilität versteht man die Genauigkeit des Messens, wie gering der Messfehler ist und damit wie nah der gemessene Wert am wahren Wert ist. Objektivität ist eine Unterkategorie der Reliabilität. Sie beschäftigt sich damit, welche Rolle der Testleiter und die Testbedingungen spielen und misst den Teil des Messfehlers, der darauf zurückzuführen ist. Validität schließlich befasst sich damit, welcher Teil der Messung auf den zu messenden Gegenstand, das Konstrukt, zurückgeführt werden kann und welche Teile auf systematische oder unsystematische Fehler zurückzuführen sind.

Die probabilistische Testtheorie wirft den Blick darauf, wie auf Grundlage von Daten oder Beobachtungen, z. B. Testergebnissen, Aussagen über Konstrukte, Vorstellungen, Ideen, also Nicht-Beobachtbares, gemacht werden können. Es wird von beobachteten Daten auf nicht-beobachtbare (latente) Persönlichkeitsvariablen geschlossen. Sie stellt damit die Interpretation der Ergebnisse in den Mittelpunkt und damit die Validität. Ein sehr bekanntes probabilistisches Modell ist das Rasch-Modell, in dem von der Schwierigkeit von Test-Items auf die Fähigkeit von Probanden geschlossen wird.

Bachman und Palmer (1996) bauen auf beiden Ansätzen und präzisieren und erweitern den Begriff der Validität. Sie beschäftigen sich mit der Bewertung sprachlicher Kompetenz und entwickeln ein testtheoretisches Modell, das sechs Gütekriterien umfasst: Reliabilität, Validität, Authentizität, Interaktivität, Rückwirkung und Ökonomie.

Unter Reliabilität verstehen sie, ähnlich wie in der klassischen Testtheorie, die Zuverlässigkeit der Ergebnisse, d. h. die Reproduzierbarkeit von Ergebnissen und ihre Unabhängigkeit von Ort, Zeit, Prüfer, Tagesform etc. Unter Validität verstehen sie Konstruktvalidität, die Angemessenheit oder Gültigkeit der Inter-

pretation der Ergebnisse aufgrund eines theoretischen Zusammenhangs zwischen den gemessenen Daten und den ihnen zugrundeliegenden Persönlichkeitsvariablen oder Fähigkeiten.

Unter Authentizität verstehen sie die Realitätsnähe der Aufgaben sowie die Interpretation der Angemessenheit oder Korrektheit der Aufgabenbewältigung im Hinblick auf lebenswirkliche Anforderungen. Unter Interaktivität verstehen sie, wie gut die Testaufgaben mit den zu messenden Fähigkeiten interagieren und nicht mit anderen Fähigkeiten, Wissenskomponenten oder emotionalen Zuständen. Sie unterscheiden vier Ebenen von Persönlichkeitsmerkmalen: die zu messende sprachliche Kompetenz, das Wissen der Person im Hinblick auf Weltwissen, Sach- und Fachwissen, die Gefühle der Person sowie die teststrategische Kompetenz der Person. Eine hohe Interaktivität ist dann gewährleistet, wenn die Testergebnisse aufgrund der sprachlichen Kompetenz variieren und nicht aufgrund von Sach- und Fachwissen, Gefühlen oder der Verwendung von Teststrategien. Sowohl Authentizität als auch Interaktivität sind damit Aspekte der Validität, da sie vor allem die Güte der Interpretation der Testergebnisse thematisieren.

Unter Rückwirkung (Engl. *impact*, *washback*) verstehen sie den Einfluss von Tests auf die Vorbereitung darauf, auf den Unterricht, auf Institutionen, auf die Gesellschaft insgesamt sowie auf die Lebenswege der Probanden. Unter Ökonomie schließlich verstehen sie die Forderung, einen vernünftigen Zusammenhang zwischen dem Aufwand und der Wichtigkeit der Entscheidungen, die aufgrund der Testergebnisse getroffen werden, herzustellen. Es sollte ein realistischer Aufwand für die Erstellung und Durchführung sowie für die Bewertung und Interpretation der Ergebnisse betrieben werden.

Aktuell interessiert man sich immer mehr dafür, was mit Testergebnissen gemacht wird, welche Konsequenzen ein Testergebnis hat und ob es sich dabei um positive Konsequenzen handelt. Bachman und Palmer (2010) gehen von einem fünfstufigen Verfahren zur Ermittlung von Konsequenzen aus, bei dem alle Entscheidungen empirisch begründbar sein müssen. Stufe 1 beschäftigt sich mit der Performanz eines Testteilnehmers aufgrund einer Aufgabenstellung. Stufe 2 befasst sich mit den Testergebnissen. Stufe 3 lenkt den Blick auf die Interpretation der Testergebnisse. Stufe 4 beschäftigt sich mit den Entscheidungen, die aufgrund der Interpretation der Testergebnisse getroffen werden.

Stufe 5 schließlich setzt sich mit den Konsequenzen auseinander, die sich für unterschiedliche Interessensgruppen aufgrund dieser Entscheidungen ergeben. Zu diesen Interessensgruppen gehören neben den Testteilnehmern und den Testern und Bewertern, die Lehrer, die zukünftigen Arbeitgeber, Stipendiengeber, Hochschullehrer etc. des jeweiligen Testteilnehmers sowie die Gesellschaft als Ganzes, da ja durch weitläufig akzeptierte Prüfungen die Wahrnehmung dessen, was wichtig im Zusammenhang mit einer Kompetenz oder Tätigkeit ist, beeinflusst wird.

Bachman und Palmer sehen die Begründung der Zusammenhänge zwischen diesen fünf Stufen als Argumentation und greifen

dabei auf das Argumentationsmodell von Stephen Toulmin zurück, das ich im Folgenden erklären möchte. Im Anschluss daran versuche ich zu zeigen, wie Bachman und Palmer dieses Modell benutzen, um die Validierung von Tests als ein Argumentationsproblem darzustellen und welche Qualitätsstandards sich daraus für die Entwicklung von Tests ergeben. Dies alles versuche ich an einem praktischen Beispiel zu erläutern.

## Toulmins Argumentationsmodell

Stephen Toulmin war ein englischer Philosoph, der beeinflusst von der Philosophie der normalen Sprache, die auf Wittgenstein zurückgeht, sich mit praktischen Argumenten oder „guten Gründen" als Grundlage für ethisches Verhalten auseinandersetzte. In seinem Argumentationsmodell wendet er sich von der klassischen Logik ab und versucht zu verstehen, wie Argumentieren in der Alltagsgesprächen funktioniert.

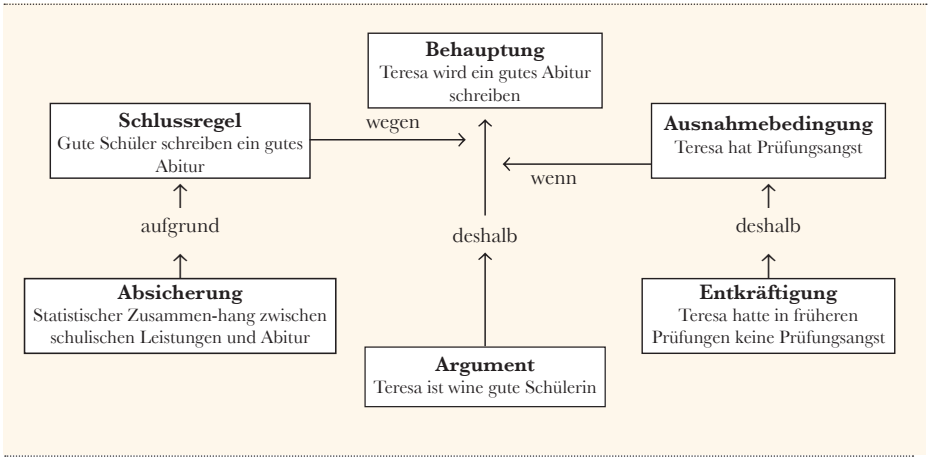
Nach Toulmin unterstützt ein praktisches Argument eine Behauptung. Wenn jemand eine Aussage macht, die hinterfragt werden kann, so ist sie zunächst einmal eine Behauptung, z. B. die Behauptung: Teresa wird ein gutes Abitur schreiben. Diese Behauptung wird durch ein Argument, das auf Fakten oder Daten beruht, unterstützt, z. B. die Tatsache, dass Teresa eine gute Schülerin ist. Nach Toulmin impliziert dies eine sogenannte Schlussregel, nämlich die allgemeine Regel, dass gute Schüler ein gutes Abitur schreiben. Diese Schlussregel kann auch explizit gemacht werden und durch weitere Aspekte abgesichert werden, z. B. dass es gesicherte statistische Zusammenhänge zwischen vorherigen schulischen Leistungen und der Leistung im Abitur gibt. Ein möglicher Einwand könnte sein, dass Teresa Prüfungsangst hat. Dies nennt Toulmin eine Ausnahmebedingung. Diese Ausnahmebedingung schließlich lässt sich durch weitere Tatsachen oder Erfahrungen absichern oder entkräften, z. B. durch den Hinweis, dass Teresa in früheren Prüfungen auch keine Prüfungsangst hatte.

Die Schlussregel ist allgemeingültig und zeigt, dass der Schritt vom Argument zur Behauptung angemessen und legitim ist und sie kann zwingend oder eingeschränkt gültig sein. Zur Absicherung werden oft Gesetze, Normen, Regeln, Prinzipien, allgemeine Erfahrungen oder anerkannte Bestimmungen oder Erfahrungen herangezogen.

Nach Toulmin heißt es also also: Eine Behauptung oder Schlussfolgerung wird durch ein Argument unterstützt, das auf Tatsachen aufbaut: Das Argument ist richtig; deshalb ist auch die Behauptung richtig. Dem ist so wegen einer allgemeinen Schlussregel oder Rechtfertigung, wenn nicht eine Ausnahmebedingung oder Gegenbehauptung auch richtig ist. Sowohl die Rechtfertigung als auch die Gegenbehauptung können nun noch durch eine Absicherung unterstützt und die Plausibilität der jeweiligen Absicherung überprüft werden.

Auf diesem Argumentationsmodell bauen Bachman und Palmer (2010) auf und behaupten, dass man Validierung als Argumentation sehen sollte, dass die Validierung eines Tests dadurch erfolgen sollte, dass man sich auf den fünf Stufen der Validierung die Behauptungen genau ansieht, sie auf die Daten zurückführt,

*Abbildung 1. Argumentationsmodell nach Stephen Toulmin (1958)*




sie rechtfertigt, Gegenbehauptungen aufstellt und diese entkräftet.

### Validierung als Argumentation

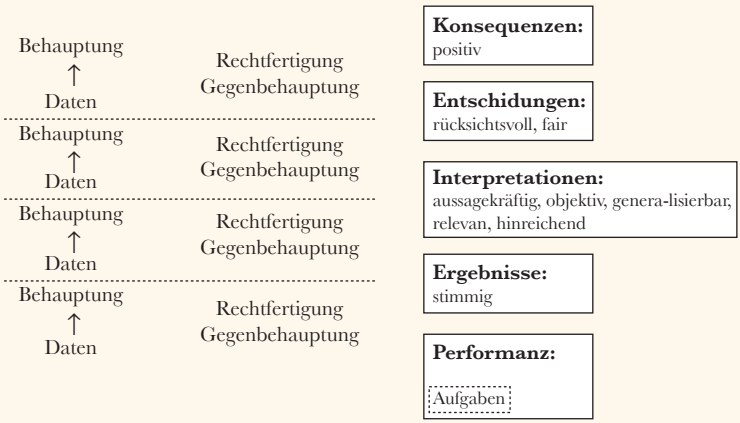
Validierung verknüpft nach Bachman und Palmer (2010) Daten und Behauptungen auf fünf Stufen oder Ebenen: die Aktivität oder Performanz eines Testteilnehmers auf Grundlage einer Aufgabenstellung; die Testergebnisse; die Interpretation der Testergebnisse; der Entscheidungen, die aufgrund der Interpretation der Testergebnisse getroffen werden; sowie die Konsequenzen, die diese Entscheidungen haben.

Auf jeder Ebene gibt es Behauptungen, die durch Daten unterstützt werden müssen, deren Zusammenhang mit der Behauptung durch Schlussregeln oder Rechtfertigungen gerechtfertigt werden muss und auf der die jeweiligen Gegenbehauptungen entkräftet werden müssen. Die Behauptung der jeweils unteren Ebene wird, wenn sie unterstützt und gerechtfertigt und die Gegenbehauptung entkräftet wurde, zu einer erfolgreichen Schlussfolgerung, die als Daten oder Fakten für die Behauptung der nächsthöheren Ebene zur Verfügung steht. Diese Behauptungen sind im Einzelnen die folgenden:

- Die Konsequenzen der Entscheidungen sind positiv.
- Die Entscheidungen aufgrund der Interpretation der Ergebnisse sind rücksichtsvoll und fair.
- Die Interpretation der Ergebnisse ist aussagekräftig, objektiv, generalisierbar, relevant und hinreichend.
- Die Ergebnisse sind stimmig.

Im Folgenden möchte ich mich nun mit diesen Behauptungen befassen und zeigen, wie diese Behauptungen durch Daten, durch Rechtfertigungen und durch die Entkräftung von Gegenbehauptungen gestützt werden müssen, um

*Abbildung 2. Validierung als Argumentation (Bachman/Palmer 2010)*




für einen Test eine umfangreiche Validierungsargumentation zu entwickeln bzw. einen Test vollständig zu validieren.

*Konsequenzen*

Mit Konsequenzen bezeichnen Bachman und Palmer (2010) das, was sie früher als *Impact* oder als *Washback* bezeichnet haben, d. h. die Rückwirkung einer bestimmten Prüfung oder Prüfungsform auf alle beteiligten Personen und Institutionen. Die Konsequenzen der Entscheidungen, die aufgrund der Interpretation der Testergebnisse getroffen wurden, sollen für alle Interessensgruppen (Engl. *stakeholders*) positiv sein. Zu diesen Interessensgruppen gehören natürlich die Testteilnehmer selbst. Prüfungen beeinflussen, wie sich Testteilnehmer auf sie vorbereiten, und sie beeinflussen, wie sie sich während der Prüfung fühlen. Die Vorbereitung sollte nicht nur relevant für die Prüfung sein. Testteilnehmer sollten durch die Vorbereitung Fähigkeiten erwerben, die für sie in ihrer Lebenswirklichkeit wichtig sind. Sie sollten sich während der Prüfung sicher fühlen, dass die Prüfung korrekt erfasst, was sie wissen und können.

Um positive Konsequenzen zu haben, ist es weiterhin wichtig, dass die Testteilnehmer ein aussagekräftiges Feedback zu ihren Ergebnissen sowie darauf, wie diese Ergebnisse zu interpretieren sind, bekommen. Testteilnehmer benötigen ein Feedback, das sich nicht nur auf die Ergebnisse der Prüfung erstreckt, sondern das konkrete Aussagen zu lebenswirklichen Fähigkeiten und Fertigkeiten macht und spezifiziert, in welchen Domänen, Spezialgebieten oder Wissensbereichen sie diese Kompetenzen aufweisen. Dieses Feedback sollte relevant, verständlich und vollständig sein. Schließlich geht es auch um die Konsequenzen der Entscheidungen, die über die Testteilnehmer aufgrund der Interpretation der Testergebnisse

getroffen werden, die ebenfalls positiv sein sollten, auch wenn den Testteilnehmern im konkreten Fall etwas verweigert wird, die Aufnahme an einer Universität zum Beispiel oder ein Stipendium. Positive Konsequenzen einer ablehnenden Mitteilung könnten zum Beispiel daraus bestehen, dass jemand daraufhin gewiesen wird, was er oder sie noch lernen muss, um ein Studium erfolgreich zu bewältigen.

Testergebnisse, bzw. die Entscheidungen, die aufgrund von Testergebnissen getroffen werden, beeinflussen aber nicht nur die Testteilnehmer, sondern auch die Lehrer, die Testteilnehmer auf den Test vorbereiten, und haben damit einen nicht unwesentlichen Anteil an der Art und Weise, wie Unterricht gestaltet wird. Schließlich beeinflussen Tests und die Entscheidungen, die aufgrund von Testergebnissen getroffen werden, auch die Gesellschaft als Ganzes. Sie haben Einfluss auf Bildungsmaßnahmen und Wertesysteme. Sie beeinflussen, was eine Gesellschaft wichtig und wertvoll findet, ja sogar, wie sich eine bestimmte Kompetenz definiert, welche ihrer Aspekte besonders wichtig und relevant sind.

Notwendige Begründungen bzw. Rechtfertigungsbelege für positive Konsequenzen sind u. a. die folgenden:

- Prüfungsergebnisse sind vertraulich und sie werden zeitnah mitgeteilt.
- Prüfungsergebnisse sowie die Interpretation dieser Ergebnisse sind für alle Interessensgruppen klar und verständlich.
- Der Einfluss auf den Unterricht und auf das Lernen ist positiv, indem Fähigkeiten und Fertigkeiten vermittelt und gelernt werden, die für die wirkliche Welt relevant sind. Dies beeinflusst natürlich Wertesysteme, indem definiert wird, was „guter" Unterricht ist, was einen kompetenten Benutzer einer Sprache auszeichnet u. Ä.
- Es wird gezeigt, dass die Konsequenzen der Entscheidungen aufgrund von Testergebnissen in der Tat für alle Interessensgruppen (Lerner, Lehrer, Arbeitgeber, Firmen, Institutionen, die Gesellschaft als Ganzes) positiv sind.

*Entscheidungen*

Die Entscheidungen, die aufgrund der Interpretation der Testergebnisse getroffen werden, sollen rücksichtsvoll und fair sein. Rücksichtsvoll bedeutet, dass sie in Einklang mit den Regeln und Werten der jeweiligen Gemeinschaft stehen und fair bedeutet, dass sie unvoreingenommen gegenüber unterschiedliche Personengruppen sind, die sich z. B. aufgrund von Geschlecht, Alter, sozialer Schicht oder Ethnizität unterscheiden.

*Interpretation*

Validität wird auch in anderen Testtheorien als die Angemessenheit der Interpretation der Ergebnisse bezeichnet. So auch hier. Die Interpretation der Ergebnisse mit Blick auf die Fähigkeiten und Fertigkeiten, die bewertet werden, sollen aussagekräftig, objektiv, generalisierbar, relevant und hinreichend sein. Aussagekraft im Hinblick auf die zu messende sprachliche Kompetenz verlangt, dass die Beschreibung der jeweiligen Kompetenz theoretisch fundiert ist, dass alle relevanten Unteraspekte dieser Kompetenz erfasst sind und dass theoretisch präzise definiert wird, welche

Interpretationen der Daten bzw. Beobachtungen zulässig sind, welche Aussagen über die zugrundeliegende, nicht beobachtbare Fähigkeit gemacht werden können.

Objektiv bedeutet, dass die Interpretation bestimmten Personengruppen gegenüber unvoreingenommen ist, so dass Variablen wie Geschlecht, Ethnizität, soziale Schicht keine Rolle spielen. Generalisierbarkeit bedeutet, dass man aufgrund der Testergebnisse Aussagen über lebensweltliche Kompetenzen, Situationen und Domänen machen kann. Relevant bedeutet, dass die Interpretationen, die geliefert werden, relevant für die zu treffenden Entscheidungen sind. Hinreichend schließlich bedeutet, dass sie genügend Hinweise und Belege für belastbare Entscheidungen liefern.

*Ergebnisse*

Die Behauptung auf der Ebene der Ergebnisse, die mit Daten gestützt und gerechtfertigt werden muss und wobei gezeigt werden muss, dass Gegenbehauptungen nicht unterstützt werden, ist die Behauptung, dass die Testergebnisse stimmig sind. Stimmig in diesem Zusammenhang bedeutet, dass die Ergebnisse einheitlich und widerspruchsfrei sind über unterschiedliche Aufgaben, über unterschiedliche Aspekte der Prüfung und über unterschiedliche Personengruppen hinweg sowie unabhängig vo dem Testzeitpunkt, der Testversion, den Testdurchführenden und etwaigen Bewertern. Diese Behauptung ist im Grunde genommen diejenige, die in der klassischen Testtheorie Reliabilität und Objektivität genannt wird. Das Neue an Bachman und Palmers (2010) Modell ist, dass Reliabilität und Objektivität nicht nur anhand der Daten bestätigt werden, sondern dass die jeweilige Rechtfertigung bzw. Schlussregel explizit gemacht und abgesichert wird sowie etwaige Gegenbehauptungen untersucht und entkräftet werden.

## Zusammenfassung

Die aktuelle Diskussion in der Testforschung fokussiert immer stärker die zentrale Bedeutung der Validität von Testergebnissen, wobei Validität nicht als eine Eigenschaft eines Tests gesehen wird, sondern als das, was mit Testergebnissen gemacht wird, die Konsequenzen, die Tests und Testergebnisse für ganz unterschiedliche Interessensgruppen haben. Ausgehend von der klassischen Testtheorie habe ich die Entwicklung der Diskussion zu Gütekriterien von Tests bis zur Jahrhundertwende kurz dargestellt und bin dann auf das zur Zeit wichtigste und umfassendste Validierungsmodell von Bachmann und Palmer (2010) eingegangen. Bachman und Palmer sehen Validierung als Argumentation und greifen auf das Argumentationsmodell von Toulmin (1958) zurück. Sie sehen Validierung als die datengestützte Begründung von Behauptungen auf fünf Ebenen, wobei die Schlussfolgerungen einer Ebene als Daten für die nächste Ebene genutzt werden. Dabei muss gezeigt werden, dass die Performanz von Testteilnehmern sich korrekt in Testergebnissen widerspiegelt, dass diese Ergebnisse im Hinblick auf die Lebenswirklichkeit der Testteilnehmer interpretiert werden können, dass auf dieser Grundlage faire Entscheidungen getroffen werden können und schließlich, dass die Konsequenzen dieser Entscheidungen positiv für alle Beteiligten sind.

## Bibliographie


Bachman, Lyle & Palmer, Adrian (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests*. Vereinigtes Königreich: Oxford University Press.

Bachman, Lyle & Palmer, Adrian. (2010) *Language Assessment in Practice: Developing Language Assessments and Justifying their Use in the Real World*. Vereinigtes Königreich: Oxford University Press.

Bond, Trevor G. & Fox, Christine M. (2013) *Applying the Rasch Model. Fundamental Measurement in the Human Sciences*. New York: Psychology Press.

Brown, James Dean. (2014) *Testing in Language Programs: A Comprehensive Guide to English Language Assessment*. New York: MacGraw-Hill.

Douglas, Dan. (2009) *Understanding Language Testing*. New York: Routledge.

Fulcher, Glenn & Davidson, Fred (2007). *Language Testing and Assessment: An Advanced Resource Book*. London: Routledge.

Kane, Michael (2006). Validation. In R. Brennan (Ed.), *Educational Measurement*, 4. Aufl. American Council on Education und Praeger Publishers. S. 17-64

Kane, Michael. (2012) Validating score interpretations and uses. In *Language Testing*, 29. S. 3-17

Toulmin, Stephen. (1958) *The Uses of Argument*. Cambridge: Cambridge University Press.