Spohn, B. (2022). Kriterien zur Charakterisierung und Beurteilung von Testverfahren zur Schulleistungsdiagnostik. Ein praxisbezogener Leitfaden. In M. Gebhardt, D. Scheer & M. Schurig (Hrsg.), *Handbuch der sonderpädagogischen Diagnostik. Grundlagen und Konzepte der Statusdiagnostik, Prozessdiagnostik und Förderplanung* (S. 871-880). Regensburg: Universitätsbibliothek. https://doi.org/10.5283/epub.53149

Kriterien zur Charakterisierung und Beurteilung von Testverfahren zur Schulleistungsdiagnostik

Ein praxisbezogener Leitfaden

Birgit Spohn

1 Einführung

Die Auseinandersetzung mit Testverfahren ist essentieller Bestandteil der Schulleistungsdiagnostik in Theorie und Praxis. In diesem Zusammenhang stellt sich die Frage, *welche* Aspekte bei der Charakterisierung und Beurteilung der Testverfahren beachtet werden sollten. Im Folgenden werden entsprechende Kriterien vorgestellt. Ziel ist es, einen Leitfaden anzubieten. (Mit den erforderlichen Modifikationen ist dieser auch für andere diagnostische Bereiche, wie zum Beispiel »Intelligenzdiagnostik« oder »Persönlichkeitsdiagnostik«, anwendbar.) Verwendet werden kann der Leitfaden zum einen im Rahmen des Studiums. Zum Beispiel, wenn im Rahmen von Seminaren, Seminar- oder Abschlussarbeiten, Prüfungen u. Ä. Testverfahren zur Schulleistungsdiagnostik charakterisiert und beurteilt werden sollen. Dies kann beispielsweise der Fall sein, wenn im Rahmen eines Seminars oder einer Prüfung ein bestimmtes Testverfahren vorgestellt und kritisch bewertet werden soll, das passende Testverfahren für die empirische Untersuchung im Rahmen einer Abschlussarbeit ausgewählt werden oder im Rahmen von Praktika eine diagnostische Untersuchung im Schulleistungsbereich geplant werden muss. Zum anderen ist ein Leitfaden im Rahmen der beruflichen Tätigkeit als Lehrperson zweckmäßig. Es empfiehlt sich hier, eine Sammlung (in Form einer Datei, eines Aktenordners o. Ä.) anzulegen, in der relevante diagnostische Verfahren beschrieben und beurteilt werden. Diese Sammlung bietet eine gute Basis für die Auswahl geeigneter Testverfahren bei beruflich bedingten Fragestellungen, wie zum Beispiel bei der diagnostischen Untersuchung im Rahmen der Überprüfung des Anspruchs auf ein sonderpädagogisches Bildungsangebot[1] oder der Diagnostik zur Feststellung des Leistungsstandes der unterrichteten Klasse oder einzelner Schüler*innen mit dem Zweck, das methodisch-didaktische Vorgehen auf Klassen-, Teilgruppen- oder Individualebene abzustimmen, Fördermaßnahmen optimal planen zu können etc.

[1] Im Text werden *exemplarisch* die in Baden-Württemberg üblichen Begrifflichkeiten verwendet, da keine erschöpfende Aufzählung für alle Bundesländer/Länder erfolgen kann.

Für den Begriff »Testverfahren zur Schulleistungsdiagnostik« (auch: Schulleistungstest) gibt es keine allgemein anerkannte Definition, wenn auch weitgehender Konsens darüber besteht, welche Testverfahren dieser Kategorie zuzuordnen sind[2]. Im Folgenden sollen hierunter in Anlehnung an Schmidt-Atzert und Amelang (2012, S. 145) Testverfahren verstanden werden, die den aktuellen Leistungsstand in Schulfächern (z. B. Mathematik) bzw. in Schulleistungsbereichen, wie dem Rechtschreiben, erfassen wollen. Das Vorkommen unterschiedlicher Definitionen bei weitgehendem Konsens bzgl. der Zuordnung von Verfahren trifft auch auf die Begriffe »Testverfahren« bzw. »Test« zu[3]. Hier wird die Definition von Döring und Bortz (2016) zugrunde gelegt.

> Ein psychologischer Test[4] […] ist ein wissenschaftliches Datenerhebungsverfahren, das aus mehreren Testaufgaben (Testbogen/Testmaterial) sowie festgelegten Regeln zu deren Anwendung und Auswertung (Testmanual) besteht. Ziel eines psychologischen Tests ist es, ein latentes psychologisches Merkmal (Konstrukt) – typischerweise eine Fähigkeit oder Persönlichkeitseigenschaft – in seiner absoluten oder relativen Ausprägung zu Forschungszwecken oder für praktische Entscheidungen zu erfassen. (S. 431)

2 Kriterien zur Charakterisierung und Beurteilung von Testverfahren zur Schulleistungsdiagnostik

Im Folgenden werden die vorgeschlagenen Kriterien vorgestellt. Tabelle 1 bietet einen knappen Überblick.

Die Kriterien erheben keinen Anspruch auf Vollständigkeit und Allgemeingültigkeit/Unangreifbarkeit, sie sollen die Leser*innen anregen, sich Gedanken über relevante Kriterien zu machen und ggf. den Leitfaden für ihre Zwecke zu modifizieren. Wie ersichtlich, sind die Kriterien sechs Kategorien zugeordnet: 1. allgemeine Angaben, 2. Durchführung / Testaufgaben / Auswertung / diagnostische Aussage/n, 3. Gütekriterien, 4. Normen, 5. praktische Erwägungen und 6. kritische Würdigung des Verfahrens.

2.1 Allgemeine Angaben

Relevante Grunddaten: Der Name des Verfahrens (incl. Abkürzung) (z. B. Hamburger Schreib-Probe 1-10; HSP 1-10; May, 2018), die Autor*innen, das Erscheinungsjahr und der Verlag. Handelt es sich nicht um die erste Auflage, sollte auch die Auflage angegeben werden. Bei einer Neuauflage ist relevant, ob es sich um eine unveränderte oder überarbeitete Auflage handelt und welche Aspekte überarbeitet wurden. Der Grad der Überarbeitung kann stark variieren. So können z. B. lediglich weitere Forschungsergebnisse zur erfassten Leistung im Manual hinzugefügt worden sein oder aber es hat z. B. eine Neunormierung stattgefunden.

[2] Im Rahmen des Beitrags soll generell keine differenzierte Diskussion einzelner Begriffe erfolgen. Auch Basiswissen kann nicht aufgearbeitet werden. Es sei hier auf die entsprechenden Beiträge des vorliegenden Herausgeberwerks bzw. Grundlagenliteratur zum Thema »Diagnostik« (wie z. B. Moosbrugger & Kelava, 2020; Schmidt-Atzert & Amelang, 2012) verwiesen.

[3] Siehe hierzu auch Schurig und Gebhardt, in diesem Band

[4] Der Begriff »*psychologischer* Test« ist hierbei durch das zu erfassende Merkmal begründet (Döring & Bortz, 2016, S. 430).

Tabelle 1: Kriterien (Kurze Übersicht)

Die Tabelle soll einen schnellen Überblick über die Hauptkriterien bieten. Aus diesem Grund deckt sich die Bezeichnung der Kriterien nicht immer akkurat mit den Bezeichnungen im Text.

1. Allgemeine Angaben
    - Relevante Grunddaten (Name des Verfahrens, Autor*innen, Erscheinungsjahr, (Auflage), Verlag
    - Erfasste Leistung/en
    - Erfasste Teilaspekte
    - Struktur des Verfahrens
    - Anzahl der Aufgaben
    - Konzept
    - Anwendungszeitraum
    - Diagnostische Zielsetzungen und Anwendungsmöglichkeiten
    - Testbestandteile
    - Kurze Klassifikation
        - Standardisiertes vs. nichtstandardisiertes Testverfahren
        - Normorientiertes vs. kriterienorientiertes Testverfahren
        - Schnelligkeits- vs. Niveautest
        - Einzel- vs. Gruppentest
        - Papier-Bleistifttest vs. computergestütztes Verfahren
    - Anwendbarkeit in sonderpädagogischen Handlungsfeldern
2. Durchführung / Testaufgaben / Auswertung / diagnostische Aussage/n
    - Durchführung: Art und Zeitdauer
    - Testaufgaben
    - Auswertung: Art, Zeitdauer und Testergebnisse
    - Diagnostische Aussage/n
3. Gütekriterien:
    - Hauptgütekriterien (Objektivität, Reliabilität und Validität)
    - Nebengütekriterien [Skalierung/Skalierbarkeit, Normierung/Eichung, (Test-) Ökonomie, Nützlichkeit, Zumutbarkeit, Unverfälschbarkeit und (Test-)Fairness]
4. Normen
5. Praktische Erwägungen
    - Kosten des Verfahrens und des Testmaterials / Vertrieb
    - ggf. Verfügbarkeit
    - Vorbereitungszeit
    - Existenz von Parallelformen
6. Kritische Würdigung des Verfahrens

Erfasste Leistung/en: Zum Beispiel die Lese- und Rechtschreibleistung.

Erfasste Teilaspekte: So erfasst z. B. ELFE II (Lenhard et al., 2020) das Leseverständnis auf Wort-, Satz- und Textebene[5]

Struktur des Verfahrens: (Dieser Aspekt steht mit dem Aspekt »Erfasste Teilaspekte« in Zusammenhang.) Gliedert sich das Verfahren in einzelne Verfahrensteile und/oder Untertests? So umfasst der Diagnostische Rechtschreibtest für 4. Klassen (DRT 4) (Grund et al., 2017) beispielsweise zwei *Verfahrensteile*, einen Lückentext, in den von der zu testenden Person nach Diktat Wörter eingetragen werden, sowie eine Fehleranalyse. Und ELFE II gliedert sich in drei *Untertests* (Wortverständnistest, Satzverständnistest und Textverständnistest).

Anzahl der Aufgaben: a) Gesamtzahl der Aufgaben, b) ggf. Anzahl der Aufgaben pro Bereich/Untertest.

Konzept, auf dem das Testverfahren basiert: D. h., auf der Basis welcher theoretischen Annahmen und welcher Forschungsergebnisse wurde das Testverfahren entwickelt? Die theoretischen Annahmen beziehen sich zum einen auf die zu erfassende Leistung (inhaltsbezogene Theorien) und zum anderen auf die Testtheorie, auf der das Testverfahren basiert, i. d. R. klassische oder probabilistische Testtheorie, (methodenbezogene Theorien) (z. B. Döring & Bortz, 2016)[6].

Anwendungszeitraum: Zeitraum, für den die Anwendung des Testverfahrens vorgesehen ist. I. d. R. findet hier ein Bezug auf Schuljahre bzw. -monate statt. So soll beispielsweise ELFE II in den letzten drei Schulmonaten der ersten Klasse bis zu den ersten drei Schulmonaten der siebten Klasse eingesetzt werden.

Diagnostische Zielsetzungen und Anwendungsmöglichkeiten des Verfahrens entsprechend der Autor*innen.

Testbestandteile: Bestandteile des Testverfahrens (Handanweisung, Aufgabenmaterial, Auswertungsbogen etc.) und gegebenenfalls erforderliche zusätzliche Materialien.

Kurze Klassifikation des Testverfahrens:

- *Standardisiertes vs. nichtstandardisiertes Testverfahren (auch formelles vs. informelles Testverfahren)*[7]: Kennzeichnend für standardisierte Testverfahren ist, dass sie »[…] wissenschaftlich entwickelt, hinsichtlich der wichtigsten Gütekriterien untersucht und unter Standardbedingungen durchführbar und normiert« sind (Lienert & Raatz, 1998, S. 14). Werden die Kriterien nicht erfüllt, so handelt es sich um nichtstandardisierte oder informelle Verfahren (Lienert & Raatz, 1998, S. 14).
- *Normorientiertes vs. kriterienorientiertes Testverfahren*: Diese Klassifikation basiert auf der verwendeten Bezugsgröße und dem Ziel der Testung (Lienert & Raatz, 1998, S. 17). »Bei normorientierten Tests wird das individuelle Testergebnis zum Populationsmittelwert in Beziehung gesetzt und ein Normwert bestimmt. Bei kriterienorientierten Tests wird das Ergebnis auf die Gesamtzahl der Aufgaben bezogen. Normorientierte Tests soll-

[5] Diese Angaben beziehen sich auf die Papierform.
Dieser und die drei folgenden Aspekte sollten möglichst kombiniert dargestellt werden.

[6] Siehe hierzu auch Schurig und Gebhardt, in diesem Band.

[7] In Bezug auf diese beiden Begriffspaare ist die Heterogenität der Definitionen größer als bei anderen Begrifflichkeiten und ihre synonyme Verwendung ist nicht allgemein anerkannt (siehe auch: Gebhardt, Scheer & Schurig, in diesem Band).

ten Proband*innen[8] möglichst gut differenzieren, kriterienorientierte Tests sollen prüfen, ob ein Kriterium (Lehrziel, Therapieziel) erreicht worden ist oder nicht.« (Lienert & Raatz, 1998, S. 17)

- *Schnelligkeits- vs. Niveautest (auch Speed- vs. Power-Test) bzw. Mischform*: Ist der Test so konzipiert, dass für das Ergebnis entscheidend ist, wie viele Aufgaben innerhalb einer begrenzten Zeit gelöst werden (Schnelligkeitstest) oder bis zu welchem Schwierigkeitsgrad Aufgaben gelöst werden können (Niveautest) (Lienert & Raatz, 1998, S. 15). Oder liegt eine Mischform vor.
- *Einzel- vs. Gruppentest*: Kann das Verfahren nur mit Einzelpersonen oder auch in Gruppen durchgeführt werden?
- *Papier-Bleistifttest vs. computergestütztes Verfahren*: Liegt der Test in Papierform vor und wird handschriftlich bearbeitet oder werden die Testitems am PC vorgegeben und bearbeitet.

Anwendbarkeit in sonderpädagogischen Handlungsfeldern und ggf. mögliche Problemfelder laut der Autor*innen.

2.2 Durchführung / Testaufgaben / Auswertung / diagnostische Aussage/n

Durchführung: *Art und Zeitdauer*

- Art der Durchführung:
  - Grobe Charakterisierung der Art der Durchführung in Bezug auf allgemeine formale Aspekte, insb. bzgl. des Grads der Standardisierung der Instruktion, des Untersuchungsmaterials und der Untersuchungssituation sowie bzgl. des Auftretens von Zeitbegrenzungen (beim Gesamttest oder einzelnen Aufgaben/gruppen).
  - Grober Ablauf der Durchführung; besondere Anforderungen an die testdurchführende Person, wie zum Beispiel erforderliche Fachkenntnisse oder mehrere parallel durchzuführende Tätigkeiten, und mögliche Probleme; Auftreten von Besonderheiten, wie zum Beispiel Auswahl der vorzugebenden Aufgaben in Abhängigkeit vom (Nicht-)Lösen der bereits vorgegebenen Aufgaben (adaptives Testen im engeren oder weiteren Sinn).
- Zeitdauer der Durchführung / ggf. reine Bearbeitungszeit für die Proband*innen laut Angaben der Autor*innen.

Testaufgaben: Aufgaben bzw. Aufgabengruppen, mit denen einzelne Aspekte erfasst werden sollen.

- Grobe Charakterisierung bzgl. formaler Gesichtspunkte (wie z. B. offene, halboffene und geschlossene Aufgaben mit den entsprechenden Subtypen; z. B. Döring & Bortz, 2016).
- Grobe Charakterisierung bezüglich inhaltlicher Gesichtspunkte. Welche Leistungen sollen konkret mit welchem Material unter welchen Bedingungen erbracht werden? Hilfreich ist es, exemplarisch einzelne Aufgaben pro Bereich bildlich darzustellen. Bildmaterial dient der Veranschaulichung und erleichtert die Beurteilung der Angemessenheit der Aufgaben für diagnostische Situationen, auch im künftigen, konkreten Anwendungsfall.

[8] Das Zitat wurde in Bezug auf die Verwendung geschlechtersensibler Sprache angepasst.

Auswertung: Art, Zeitdauer und Testergebnisse

- Art der Auswertung
  - Grobe formale Charakterisierung der Art der Auswertung (quantitativ und/oder qualitativ, ggf. bzgl. welcher Aspekte) und der Auswertungsrichtlinien bzw. -kriterien, insb. bzgl. des Grades der Standardisierung der Auswertung und möglicher Fehlerquellen.
  - Grober Ablauf der Auswertung; manuelle und/oder computergestützte Auswertung; vorhandene Hilfsmittel bei manueller Auswertung, wie Auswertungsschablonen (vgl. Testbestandteile); besondere Anforderungen an die testauswertende Person, wie zum Beispiel erforderliche Fachkenntnisse; mögliche Probleme und Besonderheiten.
- Zeitdauer der Auswertung laut Angaben der Autor*innen.
- Testergebnisse: Bei quantitativer Auswertung: erhaltene Test-/Normwerte, Möglichkeit einer Profilinterpretation und Angabe von Konfidenzintervallen. Bei qualitativer Auswertung: Art und Differenziertheit der Aussagen.

Diagnostische Aussage/n, die entsprechend der Autor*innen aufgrund der Testergebnisse getroffen werden können.

2.3 Gütekriterien

Hauptgütekriterien (vertiefend: Moosbrugger & Kelava, 2020): Angaben zu den Hauptgütekriterien durch die Testautor*innen.

- Objektivität

  »Ein Test ist dann objektiv, wenn das ganze Verfahren, bestehend aus Testmaterialien, Testdarbietung, Testauswertung und Interpretationsregeln, so genau festgelegt ist, dass der Test unabhängig von Ort, Zeit, Testleiter*in[9] und Auswerter*in durchgeführt werden könnte und für eine bestimmte Testperson bezüglich des untersuchten Merkmals dennoch dasselbe Ergebnis und dieselbe Ergebnisinterpretation liefert« (Moosbrugger & Kelava, 2020, S. 18).

- Reliabilität/Zuverlässigkeit

  »Ein Test erfüllt das Gütekriterium der Reliabilität/Zuverlässigkeit, wenn er das Merkmal, das er misst, exakt, d. h. ohne Messfehler, misst« (Moosbrugger & Kelava, 2020, S. 27).

- Validität/Gültigkeit

  »Validität/Gültigkeit eines Tests liegt vor, wenn der Test das Merkmal, das er messen soll, auch wirklich misst und nicht irgendein anderes« (Moosbrugger & Kelava, 2020, S. 30).

Bzw. entsprechend der aktuell vorherrschenden Konzeption des Begriffs: »Validität ist das Ausmaß, in dem empirische Befunde und theoretische Argumente die Interpretationen von Test-

[9]Das Zitat wurde in Bezug auf die Verwendung geschlechtersensibler Sprache angepasst.

werten für die beabsichtigten Verwendungen von Tests unterstützen« (AERA et al., 2014, S. 11, übersetzt von und zit. nach Hartig et al., 2020, S. 530).

Nebengütekriterien (vertiefend: u.a. Döring & Bortz, 2016; Moosbrugger & Kelava, 2020): Angaben zu den Nebengütekriterien durch die Testautor*innen.

- Skalierung/Skalierbarkeit

  »Ein Test erfüllt das Gütekriterium der Skalierung, wenn die laut Verrechnungsregel resultierenden Testwerte (numerisches Relativ) die tatsächlichen Merkmalsrelationen (empirisches Relativ) adäquat abbilden« (Moosbrugger & Kelava, 2020, S. 20).

- Normierung/Eichung

  »Ein Test gilt als normiert (geeicht), wenn für ihn ein Bezugssystem erstellt wurde, mit dessen Hilfe die Ergebnisse einer Testperson im Vergleich zu den Merkmalsausprägungen anderer Personen der Zielgruppe eindeutig eingeordnet und interpretiert werden können« (Moosbrugger & Kelava, 2020, S. 22).

- (Test-)Ökonomie

  »Ein Test erfüllt das Gütekriterium der Ökonomie, wenn er – gemessen am diagnostischen Erkenntnisgewinn – wenig finanzielle und zeitliche Ressourcen beansprucht« (Moosbrugger & Kelava, 2020, S. 24).

- Nützlichkeit

  »Das Gütekriterium der Nützlichkeit eines Tests ist gegeben, wenn das von ihm gemessene Merkmal praktische Relevanz aufweist und die auf seiner Grundlage getroffenen Entscheidungen (Maßnahmen) mehr Nutzen als Schaden erwarten lassen« (Moosbrugger & Kelava, 2020, S. 24).

- Zumutbarkeit

  »Ein Test erfüllt das Kriterium der Zumutbarkeit, wenn er hinsichtlich des aus seiner Anwendung resultierenden Nutzens die Testpersonen in zeitlicher, psychischer sowie körperlicher Hinsicht nicht über Gebühr belastet« (Moosbrugger & Kelava, 2020, S. 25).

- Unverfälschbarkeit

  »Ein Testverfahren erfüllt das Gütekriterium der Unverfälschbarkeit, wenn das Verfahren derart konstruiert ist, dass die Testperson die konkreten Ausprägungen ihrer Testwerte durch gezielte Vortäuschung eines für sie unzutreffenden Testverhaltens nicht verzerren kann« (Moosbrugger & Kelava, 2020, S. 26).

- (Test-)Fairness[10]

  »Ein Test erfüllt das Gütekriterium der Fairness, wenn die resultierenden Testwerte zu keiner systematischen Benachteiligung bestimmter Personen aufgrund ihrer Zugehörigkeit zu ethnischen, soziokulturellen oder geschlechtsspezifischen Gruppen führen« (Moosbrugger & Kelava, 2020, S. 25).

[10] Die Testfairness hat in den Standards for Educational and Psychological Testing (AERA et al., 2014) als Beurteilungsmerkmal von Testverfahren stark an Bedeutung gewonnen (z. B. Döring & Bortz, 2016).

2.4 Normen

Normen (Art, Differenziertheit, Normen für Teilgruppen, Grad der Differenzierung im unteren (und oberen) Leistungsbereich etc.). Bezüglich des Prozesses der Normierung sind primär Charakteristika der Normierungsstichprobe/n (insb. Art, Stichprobengröße (auch bzgl. Teilstichproben) und Repräsentativität – für welche Gruppen und bezüglich welcher Aspekte – und somit die entsprechende Repräsentativität der Normen) und der Zeitpunkt der Normierung (und somit die Aktualität der Normen[11]) von Interesse.

2.5 Praktische Erwägungen

Kosten des Verfahrens und des Testmaterials (Testhefte, Auswertungsbogen etc.) sowie Vertrieb

ggf. Verfügbarkeit: Ist das Verfahren an der Hochschule entleihbar, am Arbeitsplatz vorhanden, in eigenem Besitz o.Ä.?

Einarbeitungszeit: Zeit, die für das Einarbeiten in das Verfahren investiert werden muss. (Dieser Aspekt ist insbesondere bei erforderlicher zeitnaher Testdurchführung relevant.)

Existenz von Parallelformen[12]

Etc.

2.6 Kritische Würdigung des Verfahrens

Auf der Basis der Auseinandersetzung mit dem Testverfahren erfolgt eine kritische Würdigung, d. h., das Auflisten und Abwägen von Vor- und Nachteilen des Verfahrens unter besonderer Berücksichtigung der Anwendbarkeit im sonderpädagogischen Bereich. Dies sollte unter Bezug auf die Hauptfragstellungen/-zielsetzungen im Rahmen der sonderpädagogischen Diagnostik erfolgen (grob: Feststellungs- vs. Förderdiagnostik entsprechend der Kultusministerkonferenz, 2019; Gebhardt, Scheer & Schurig, in diesem Band), da sich Testverfahren zumeist für unterschiedliche Fragestellungen in unterschiedlichem Ausmaß eignen.

Erste Anhaltspunkte können hier folgende Aspekte bieten. Die Angemessenheit des Konzepts, auf dem das Verfahren basiert, d. h., wie gut das Konzept den aktuellen Stand der Theoriebildung und der Forschung abbildet. Die Art und Differenziertheit der Testergebnisse und der diagnostischen Aussagen (z. B. Förderrelevanz). Das Ausmaß, in dem die Hauptgütekriterien erfüllt werden[13]. Hier sollte die Aussagekraft der Untersuchungsergebnisse eingeschätzt werden. Bei der Validität (auch) bezogen auf einzelne diagnostische Schlussfolgerungen (AERA et al., 2014). Und die Qualität der Normen (insb. Aktualität, Differenziertheit, Repräsentativität und Grad der Differenzierung im unteren Leistungsbereich[14]) sowie das Vorliegen von Normwerten für Schüler*innen mit Anspruch auf ein sonderpädagogisches Bildungsangebot/Beeinträchtigungen.

[11] Bei Leistungstests können aufgrund von Kohorteneffekten Normen nur zeitlich begrenzt Gültigkeit beanspruchen. Als Richtwert werden oft *maximal* 10 Jahre genannt (z. B. Macha et al., 2006).

[12] Die Existenz von Parallelformen war im traditionellen Nebengütekriterium »Vergleichbarkeit« inkludiert (Lienert & Raatz, 1998).

[13] Zu Kritik an den Gütekriterien: Bundschuh und Winkler (2019).

[14] Siehe hierzu auch Renner, in diesem Band.

Weitere Anhaltspunkte liefert das Ausmaß, in dem die Nebengütekriterien erfüllt werden. Praktische Erwägungen, die sich z. B. auf die Durchführungsdauer bzw. die Bearbeitungszeit, die Durchführbarkeit in Gruppen und das Vorhandensein von Parallelformen beziehen, werden/wurden auch in den Nebengütekriterien abgebildet. Insbesondere aber ist relevant, ob das Verfahren bei bestimmten Personen(gruppen) nicht einsetzbar ist bzw. in welchem Ausmaß systematische Benachteiligung/Einschränkungen der Validität der diagnostischen Aussagen (Testfairness) und/oder eine Belastung durch die Testung (Zumutbarkeit) vorliegen. Die Einschränkungen der Anwendbarkeit können durch Beeinträchtigungen in Bezug auf die Hör- und Sehfähigkeit, die Wahrnehmung, die (Fein-)Motorik, das Sprechen, die sprachlichen und kommunikativen Fähigkeiten, den sozio-emotionalen Bereich, die kognitiven Lernvoraussetzungen (wie z. B. Gedächtnisleistungen) und das Lern- und Arbeitsverhalten (wie z. B. Motivations- und Konzentrationsprobleme) etc. oder z. B. auch durch den biographischen (z. B. traumatische Erfahrungen durch Flucht o.Ä.) oder sozioökonomischen/soziokulturellen Hintergrund begründet sein. (Wobei jeweils das Ausmaß der Beeinträchtigungen in Betracht gezogen werden muss.) Die Gestaltung des Verfahrens in Bezug auf die Anforderungen an die Hör- und Sehfähigkeit, die motorischen und sprachlichen Anforderungen etc., aber z. B. auch die inhaltliche Ausgestaltung, das Ausmaß der altersadäquaten und motivierenden Gestaltung des Testmaterials, der Instruktion und der Durchführungsbedingungen, die Durchführungsdauer und möglicher Zeitdruck beim Bearbeiten können diesbezüglich Anhaltspunkte bieten.

3 Literatur

American Educational Research Association (AERA), American Psychological Association (APA) & National Council on Measurement in Education (NCME) (2014). *Standards for educational and psychological testing* (6th ed.). AERA, APA, NCME.

Bundschuh, K. & Winkler, C. (2019). *Einführung in die sonderpädagogische Diagnostik*. Ernst Reinhardt.

Döring, N. & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5., vollst. überarb., aktual. u. erw. Aufl.). Springer. DOI 10.1007/978-3-642-41089-5

Grund, M., Leonhart, R. & Naumann, C.L. (2017). *Diagnostischer Rechtschreibtest für 4. Klassen. DRT 4* (3., aktual. u. neu normierte Aufl.). Hogrefe.

Hartig, J., Frey, A. & Jude, N. (2020). Validität von Testwertinterpretationen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (3., vollst. neu bearb., erw. u. aktual. Aufl., S. 529-544). Springer. https://doi.org/10.1007/978-3-662-61532-4

Kultusministerkonferenz (KMK) (2019). *Empfehlungen zur schulischen Bildung, Beratung und Unterstützung von Kindern und Jugendlichen im sonderpädagogischen Schwerpunkt LERNEN*. https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2019/2019_03_14-FS-Lernen.pdf

Lenhard, W., Lenhard, A. & Schneider, W. (2020). *ELFE II. Ein Leseverständnistest für Erst- bis Siebtklässler – Version II* (4., unveränd. Aufl.). Hogrefe.

Lienert, G.A. & Raatz, U. (1998). *Testaufbau und Testanalyse* (6. Aufl.). Beltz.

Macha, T., Proske, A. & Petermann, F. (2006). Validität von Entwicklungstests. *Kindheit und Entwicklung*, *14*, (3), 150-162. https://doi.org/10.1026/0942-5403.14.3.150

May, P. (2018). *Hamburger Schreib-Probe 1-10 (HSP 1-10)*. Verlag für pädagogische Medien.

Moosbrugger, H. & Kelava, A. (2020). Qualitätsanforderungen an Tests und Fragebogen (»Gütekriterien«). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (3., vollst. neu bearb., erw. u.aktual. Aufl., S. 13-38). Springer. https://doi.org/10.1007/978-3-662-61532-4

Schmidt-Atzert, L. & Amelang, M. (2012). *Psychologische Diagnostik* (5., vollst. überarb. u. erw. Aufl.). Springer. https://doi.org/10.1007/978-3-642-17001-0

Birgit Spohn ist Diplom-Psychologin (Studienschwerpunkte: Pädagogische Psychologie und Klinische Psychologie). Sie arbeitet in der Funktion einer Akademischen Rätin an der Fakultät für Sonderpädagogik der Pädagogischen Hochschule Ludwigsburg in der Abteilung Förderschwerpunkt Lernen. Ihre Arbeitsschwerpunkte sind Psychologie und Diagnostik. Zudem ist sie Leiterin der Testsammlung Sonderpädagogik der Pädagogischen Hochschule Ludwigsburg.