Prof. Dr. Anne Levin

EINFÜHRUNG IN DIE DIAGNOSTIK

GRUNDLAGENTEXT FÜR DEN MASTER OF EDUCATION
GYMNASIUM / OBERSCHULE IM MODUL EWL GO 3

ADiEU
Allgemeine Didaktik
und Empirische
Unterrichtsforschung

Inhalt

GRUNDLAGEN DER DIAGNOSTIK

Stichwörter für dieses Kapitel: Pädagogische Diagnostik, Psychologische Diagnostik, Eigenschaftsdiagnostik, Verhaltensdiagnostik, Modifikation, Selektion, Prozessdiagnostik, Ergebnisdiagnostik, Zuschreibungen

Zunächst ist zu klären, was überhaupt unter Psychologischer Diagnostik verstanden wird und wo gegebenenfalls Unterschiede zwischen pädagogischer und Psychologischer Diagnostik liegen.

Psychologische Diagnostik hat das Ziel, psychologisch bedeutsame Charakteristika von Personen, Gruppen, Institutionen und Situationen zu erfassen. Dazu nutzt sie ein System von Regeln, Anleitungen und Instrumenten. Die gewonnenen Daten sollen nachfolgend analysiert werden, um zu einem diagnostischen Urteil zu gelangen, dessen Ziel wiederum vielfältig sein kann. So kann ein Ziel sein, eine Entscheidung bezogen auf mögliche Interventionen oder Therapiemaßnahmen zu treffen, ein weiteres, eine Intervention zu evaluieren oder aber eine Prognose für zukünftiges Verhalten bzw. die weitere Entwicklung abzugeben.

Psychologische Diagnostik kann daher im Kontext von Therapie, Beratung, Behandlung oder Bewertung erfolgen. Letztere finden wir im Zusammenhang mit der Evaluation von Bildungseinrichtungen im Zuge von großen Schulstudien.

Die **Psychologische Diagnostik** kann wie folgt definiert werden (Amelang & Schmidt-Atzert, 2006, S. 3):

„Psychodiagnostik ist eine Methodenlehre im Dienste der Angewandten Psychologie. Soweit Menschen die Merkmalsträger sind, besteht ihre Aufgabe darin, interindividuelle Unterschiede im Verhalten und Erleben sowie intraindividuelle Merkmale und Veränderungen einschließlich ihrer jeweils relevanten Bedingungen so zu erfassen, dass hinlänglich präzise Vorhersagen künftigen Verhaltens und Erlebens sowie deren evtl. Veränderungen in definierten Situationen möglich werden."

Der Fokus der Psychologischen Diagnostik kann also auf der Betrachtung von Unterschieden *zwischen* den Personen (interindividuell) oder *in der Person selbst* (z.B. im Entwicklungsverlauf) liegen. Sie nutzt in der Regel standardisierte Instrumente und Klassifikationssysteme zur Einordnung, analysiert aber ebenso

wie die Pädagogische Diagnostik die Interaktion von internen und externen Einflussfaktoren und deren Bedeutung für das gezeigte Verhalten.

In Abgrenzung dazu befasst sich die **Pädagogische Diagnostik** zunächst mit einem spezifischen Bereich: Ihr Fokus liegt auf der Beschäftigung mit einzelnen Lernenden oder Gruppen von Lernenden und sie untersucht die Voraussetzungen und Bedingungen planmäßiger Lehr- und Lernprozesse, mit dem Ziel der Optimierung des Lernens. In diesem Zusammenhang befasst sie sich auch mit Fragen der Zuweisung aufgrund diagnostizierter Eignung (Selektion) und der individuellen Förderung im Rahmen von spezifischen Förderprogrammen (vergl. Ingenkamp & Lissmann; 2008).

Ob die Evaluation und Steuerung von bildungspolitischen Entscheidungen und Systemen in den Bereich der Pädagogischen Diagnostik fällt, ist umstritten. Ingenkamp und Lissmann (2008) argumentieren, dass hier die Personen selbst nicht interessieren, sondern diese nur als Merkmalsträger genutzt werden, was der Intention der Pädagogischen Diagnostik widerspräche. Dagegen könnte man argumentieren, dass die Untersuchung von Voraussetzungen und Bedingungen für erfolgreiches Lernen ja durchaus der Pädagogischen Diagnostik zugeordnet wird und sich häufig auf Personengruppen und nicht einzelne Individuen bezieht. Die Evaluation von Bildungssystemen hat auf einer Metaebene auch das Ziel Gelingensbedingungen zu identifizieren, was letztlich eben auch für die einzelnen Personen bedeutsam ist, insofern interessieren diese sehr wohl.

Im Folgenden werden unterschiedliche Ansätze und Modelle der Diagnostik vorgestellt. Diese zeigen letztlich mit welchem Blick und welchem Ziel diagnostiziert wird.

Eigenschaft oder Verhalten?

Eine Kontroverse hat die Psychologie über viele Jahre begleitet: Die Frage, ob sich das Verhalten von Menschen aufgrund von stabilen Persönlichkeitsmerkmalen, Eigenschaften oder Verhaltensdispositionen erklären und vorhersagen lässt. Wenn dem so wäre, dann hätte überspitzt gesagt, die Pädagogik wenig zu tun, sie müsste den sich entwickelnden Eigenschaften und Persönlichkeiten Raum geben, ohne diese maßgeblich beeinflussen zu müssen. Bereits daraus wird ersichtlich, dass die Pädagogik einen anderen Ansatz verfolgt. Sie geht davon aus, dass es zwar relativ überdauernde Dispositionen oder Merkmale gibt, diese aber grundsätzlich beeinflussbar und damit auch veränderbar sind.

Beispiel für eine Eigenschaftsdiagnostik

Stellen Sie sich folgende Situation vor:

Eine Schülerin der fünften Klasse wird aufgrund ihrer sprachlichen Probleme mit einem Sprachentwicklungstest getestet. Der Test zeigt, dass die Leistung der Schülerin deutlich unter dem für ihr Alter zu erwartenden Niveau bewegt. Es wird daher eine Sprachentwicklungsverzögerung diagnostiziert.

Die Feststellung einer Sprachentwicklungsverzögerung im oben genannten Beispiel ist zunächst die Zuweisung einer über einen längeren Zeitraum vorliegenden und wahrscheinlich auch länger anhaltenden Beeinträchtigung. Die Frage, die sich an eine solche Diagnose anschließen muss, ist, was man nun aus dieser Diagnose ableiten kann?

Der Prozess der Diagnostik im Fallbeispiel oben ist also nicht abgeschlossen. Im Anschluss an diese Diagnose ergeben sich mindestens zwei weitere Aufgaben: Einerseits wäre zu prüfen, warum dieses Defizit vorliegt, andererseits woraus sich dann im besten Fall entsprechende Konsequenzen hinsichtlich der Förderung des Kindes ableiten lassen.

Im Rahmen der ersten Frage wären zunächst ganz unterschiedliche Erklärungsansätze denkbar, die die Verzögerung der Sprachentwicklung auf *stabile Merkmale* oder *Eigenschaften* wie z.B. genetische Ursachen oder aber auf *variable Merkmale* wie z.B. Umgebungs- oder Lebensbedingungen, aber auch auf das *Verhalten* des Kindes selbst zurückführen. Selbstverständlich wären auch Erklärungsansätze möglich, die verschiedene Ursachen gleichermaßen benennen. So könnte das Kind in einer ungünstigen Lebenslage aufwachsen, frühe Fluchterfahrungen haben und gleichzeitig durch eine Häufung von Sprachentwicklungsverzögerungen in der Familie eine genetische Disposition aufweisen. Weiter könnte es in Folge bestimmte Sprachvermeidungsstrategien entwickelt haben, sodass es sich im Alltag weitgehend „sprachfrei“ bewegen kann.

Die Diagnose „Sprachentwicklungsverzögerung“ sagt uns also zunächst nur, dass eine Einschränkung vorliegt, die wahrscheinlich nicht ohne Fördermaßnahmen behoben werden kann und auch mit großer Wahrscheinlichkeit über einen

längeren Zeitraum anhält. Ob wir dieses allerdings als stabile und unveränderbare Eigenschaft oder aber als der Veränderung zugängliches Merkmal interpretieren, bleibt zunächst noch offen.

Letztlich hängt also das Gelingen der sich ableitenden Förderung davon ab, inwieweit die Diagnostik eine breite Ursachenforschung betrieben hat. Schränkt sie sich von vorne herein ein, indem sie ausschließlich *Eigenschaften* diagnostiziert oder sich nur auf eine *Analyse des Verhaltens* des Kindes beschränkt, ist die Wahrscheinlichkeit groß, dass die darauf abzielenden Fördermaßnahmen nicht oder nur eingeschränkt greifen, weil sie blinde Flecken aufweisen. So wird eine Fokussierung auf die Veränderung des Verhaltens dann nicht greifen, wenn gleichzeitig ignoriert wird, dass in der Familie gehäuft das Problem auftritt und entsprechend, das Kind in einem Setting aufwächst, das möglicherweise nur eingeschränkte Ressourcen zur Unterstützung des Kindes bereithält (vergl. Bourdieu, 1992). Andersherum würde eine reine Zuschreibung der Eigenschaft „sprachentwicklungsverzögert“ mit der Annahme einhergehen, dass das Verhalten des Kindes gar nicht veränderbar ist, weil es ja Ausdruck der genetisch bedingten Eigenschaft ist, dazu führen, dass lediglich Maßnahmen getroffen werden, das Kind bestimmten Einrichtungen zuzuführen, die für ein Kind mit dieser Diagnose passend sind, oder Bedingungen innerhalb der Lerngruppe zu modifizieren, was ebenfalls zu einer Stigmatisierung führen könnte. Das Kind bekäme in beiden Fällen also ein Label zugewiesen, das kaum mehr veränderbar wäre.

Dennoch kann die Messung bestimmter Eigenschaften sinnvoll sein. Betrachtet man zum Beispiel das Konstrukt „Intelligenz“, so wird davon ausgegangen, dass dieses über die Lebensspanne stabil ist und es innerhalb der Population normalverteilt ist. Das bedeutet, dass es einen Mittelwert gibt, um den sich herum die Werte symmetrisch derart verteilen, dass 50% der Population darüber und 50 % darunterliegen. Die Abweichung vom Mittelwert wird in sogenannten Standardabweichungen angegeben. Für das Konstrukt der Intelligenz ist ein Mittelwert von 100 IQ-Punkten festgelegt worden. Innerhalb einer Standardabweichung, also zwischen den Intelligenzquotienten 85 und 115 befinden sich 68 % der Population. Oder anders ausgedrückt 68% der Menschen haben einen IQ, der zwischen 85 und 115 liegt.

Geht man zwei Standardabweichungen nach oben und nach unten, so schließt man bereits 95 % der Population ein. Oder auch hier anders gesagt: 95 % der

Menschen haben einen IQ zwischen 70 und 130. Was ist nun mit den verbleibenden 5 %?

2,5 % der Menschen haben einen IQ der geringer ist als 70 IQ-Punkte. Diese Menschen werden als geistig behindert eingestuft, weil der Wert 70 die gesetzte Grenze ist, anhand derer entschieden wird, ob eine Person als geistig behindert diagnostiziert wird oder nicht.

Andersherum werden die verbleibenden 2,5 % der Menschen, die einen IQ haben, der über 130 liegt, als hochbegabt bezeichnet. Inwiefern ist die Zuweisung einer solchen Eigenschaft, die nachweislich stabil ist, überhaupt für die Pädagogik relevant?

Zunächst ist die Feststellung, dass es ab einem bestimmten Alter ein stabiles Merkmal wie Intelligenz gibt, von der Frage zu trennen, ob diese genetisch determiniert ist oder auch durch Sozialisation und Lernen beeinflusst wird. Aktuelle Studien belegen, dass Intelligenz genetisch verankert ist, deren Entfaltung jedoch maßgeblich von der Existenz einer geistig anregenden Umwelt abhängt (vergl. Stern & Neubauer, 2016). Insofern haben wir es mit einem sich stabilisierenden Merkmal zu tun, dessen Entfaltung allerdings von den Umgebungsbedingungen abhängt.

Welchen Sinn könnte es ergeben, dieses Merkmal zu erfassen?

Einerseits sind Intelligenztests da angeraten, wo abgeklärt werden soll, ob z.B. eine Lernstörung in bestimmten Bereichen oder aber eine geistige Behinderung (z.B. auch zusätzlich zu einer anderen Diagnose) vorliegt[1]. Auch die internationalen Klassifikationssysteme differenzieren inzwischen deutlich stärker zwischen den festgestellten Defiziten einerseits und den relevanten Fähigkeiten der Personen andererseits. Zudem wird Behinderung im Sinne des ‚durch die Umgebungsbedingungen behindert werden' diskutiert und der Fokus stärker auf die zu verwirklichenden Möglichkeiten der Teilhabe gelegt (vergl. Deutsches Institut für Medizinische Dokumentation und Information (DIMDI), 2005). Auch im Bereich der Hochbegabung ist eine Testung deshalb sinnvoll, weil es Kinder und Jugendliche gibt, die Höchstleitungen vollbringen, ohne hochbegabt zu sein. Sie wären sogenannte Overachiever. Eine weitere Förderung unter der falschen Annahme, dass diese Kinder oder Jugendlichen hochbegabt seien, könnte

[1] Die Begriffe Störung, Behinderung etc. werden immer wieder kritisch diskutiert. Katzenbach und Schröder (2007) setzen sich mit den Folgen der Feststellung von Gleichheit und Differenz im Bildungssystem auseinander.

allerdings dazu führen, dass man sie unnötig unter Druck setzt. Schließlich leisten Sie schon gemessen an ihrem Potential Überdurchschnittliches. Andersherum könnte die Diagnose Hochbegabung auch interessant sein, wo das Potential unterdurchschnittlich genutzt wird. Hier wäre also zu prüfen, warum der Schüler oder die Schülerin das vorhandene Potential nicht nutzt. Im Anschluss würden wir also wiederum das Verhalten des Schülers oder der Schülerin bzw. die Bedingungen, unter denen sich das Verhalten zeigt diagnostizieren.

Letztlich kann man feststellen, dass sowohl Eigenschafts- als auch Verhaltensdiagnostik ihren Platz in der Pädagogischen Diagnostik haben.

Im Unterrichtsalltag finden wir häufig Zuschreibungen von Lehrkräften, denen zwar häufig gar keine fundierte Diagnostik zugrunde liegt, die aber dennoch Eigenschaften zuschreibt oder Verhalten beurteilt und einschätzt.

Zuschreibungen im Schulalltag

Beispiel I

Die Rückgabe einer wenig gelungenen Englischarbeit an Viktor wird von der Lehrerin mit den Worten „Also Sprachen liegen dir ja wohl überhaupt nicht!“ begleitet.

Sie weist ihm also hier die Eigenschaft „fehlendes Sprachtalent“ im Sinne einer stabilen Persönlichkeitsdisposition zu. Ihre Diagnose leitet sie aus wahrscheinlich mehreren missglückten Arbeiten und einem allgemeinen Eindruck im Unterricht ab. Diagnostisch gesehen, ist dies unzureichend. Erstens müsste für die Zuweisung der Eigenschaft eine objektive Testung erfolgen mit dafür geeigneten standardisierten Testverfahren. Zweitens fehlt die Frage nach anderen Erklärungsansätzen für das erkannte schulische Problem. Ist die Lernumgebung für Viktor geeignet (Passung der Aufgaben, Beziehung Lehrerin-Schüler)? Wie lernt Viktor? Ist er überhaupt motiviert (Fragen nach dem Selbstkonzept, der Anstrengungsbereitschaft etc.)?

Beispiel II

Lisa, eine Schülerin der siebten Klasse, kann sich nicht lange auf einen Gegenstand konzentrieren und beginnt nach einer gewissen Zeit während der Stillarbeit damit, auf dem Stuhl unruhig hin und her zu rutschen. Aus dem Fenster zu schauen oder sich von anderen Dingen ablenken zu lassen.

Der Lehrer ermahnt sie deshalb häufig, was allerdings nicht zu einem Erfolg führt. In einem Elterngespräch, teilt er den Eltern mit, dass ihre Tochter ihr Arbeitsverhalten dringend ändern muss. Lisa würde auch andere Mitschüler durch ihr Verhalten stören. Sollte sie dieses Verhalten nicht ändern, dann müsse er zu Konsequenzen greifen und Lisa in solchen Fällen in den Auszeitraum schicken.

Obwohl man also erkennen kann, dass die Aussage der Lehrerin über Viktor im ersten Beispiel mit großer Wahrscheinlichkeit zu kurz greift, finden wir solche Zuschreibungen in der Schule immer wieder vor. Eigenschaftszuschreibungen reduzieren Komplexität: Wenn wir sagen, dass eine Person einfach ‚so ist', dann entlässt uns eine solche Zuschreibung aus der Verantwortung. Wir müssen und können jetzt ja gar nichts mehr tun, weil die Eigenschaft stabil ist. Im ersten Beispiel ist ‚*Viktor halt nicht begabt'*, da ‚*kann man als Lehrkraft auch nichts mehr machen'*. Auch im zweiten Beispiel geben wir die Verantwortung ab. Lisa soll ihr Verhalten ändern und ist daher selbst dafür verantwortlich. Wir unterstellen, dass es eine Frage des Willens ist, ob sie das tut oder nicht. Im zweiten Fall müssten wir eigentlich untersuchen, ob Lisa überhaupt in der Lage ist, sich über den geforderten Zeitraum hinweg zu konzentrieren.

In beiden Fällen ziehen sich die Lehrkräfte auf eine Position zurück, in der sie selbst kaum mehr handeln müssen oder, wenn überhaupt, dann eher strafend (durch schlechte Noten oder durch Sanktionen). Wenn wir uns so verhalten, entziehen wir uns letztlich unserer pädagogischen Verantwortung, die in beiden Beispielen darin bestehen müsste, genauer zu diagnostizieren, *warum* Viktor Schwierigkeiten im Fach Englisch hat und *warum* Lisa sich nur über einen begrenzten Zeitraum hinweg konzentrieren kann.

Selektion oder Modifikation?

Grundsätzlich lassen sich zunächst **zwei Strategien der Diagnostik** (vergl. Pawlik, 1976; Wild & Krapp, 2006) unterscheiden: Die institutionelle und die individuelle Diagnostik. Im Rahmen der institutionellen Diagnostik sieht sich die Diagnostik grundsätzlich wiederkehrenden, gleichbleibenden (auch bildungspolitischen) Fragestellungen ausgesetzt. Die individuelle Diagnostik befasst sich mit sehr speziellen Problemstellungen und ist kennzeichnend für die klinische Diagnostik. Dennoch finden wir in der Pädagogischen Diagnostik beide Arten vor: Im Rahmen der *Personenselektion* werden geeignete Personen für spezifische Anforderungen gesucht. Ein Beispiel wäre, wenn die Schule einen Schüler oder eine Schülerin auswählt, der oder die am besten geeignet scheint, einen Vorlesewettbewerb zu gewinnen. Bei der *Bedingungsselektion* ist es genau andersherum: Hier wird eine Bedingung ausgesucht, die möglichst gut geeignet für die Person zu sein scheint. Dies wäre z.B. der Fall, wenn im Rahmen von Praktika unter bestimmten Betrieben, derjenige ausgewählt wird, welcher möglichst gute Bedingungen für die Entwicklung eines Schülers oder einer Schülerin bietet.

In der Schule überwiegt die Personenselektion da, wo die Bildungsangebote wenig individualisiert sind. Deshalb kann mit dem Einsatz flexiblerer didaktischer Konzepte die Bedingungsselektion stärker in den Vordergrund rücken als in didaktischen Umgebungen, die stark auf Frontalunterricht und geringe Differenzierung ausgerichtet sind.

Ein weiterer Aspekt wird hier sichtbar. Wenn möglichst günstige Bedingungen für eine Person ausgewählt werden sollen, mit dem Ziel diese in ihrer Entwicklung zu unterstützen, dann ist gerade in der Schule im Rahmen des Unterrichts auch eine Bedingungsmodifikation möglich. Es geht hier also gar nicht nur darum, auszuwählen, was besonders gut zu sein scheint, sondern möglicherweise sogar darüber hinaus Bedingungen zu entwickeln oder zu modifizieren, so dass sie passend erscheinen und die Entwicklung begünstigen. Dies wiederum ist eine der beiden Möglichkeiten der Modifikationsdiagnostik, deren Ziel es ist, durch Interventionen die Entwicklung (in der Schule zumeist des Schülers oder der Schülerin) zu fördern. Dies kann einerseits dadurch erreicht werden, dass durch eine Veränderung des Verhaltens des Schülers oder der Schülerin bestimmte Ziele erreicht werden. So könnte ein Training im Bereich von Lernstrategien dazu führen, dass sich ein Schüler besser auf Klausuren vorbereiten kann und entsprechend erfolgreicher abschneidet. Andererseits könnten aber auch didaktische Maßnahmen (Erstellung von bestimmten Materialien für den Schüler)

ergriffen werden, die dazu führen, dass besser und nachhaltiger gelernt wird und dadurch das gleiche Ziel erreicht wird. Dieses wäre dann eine Bedingungsmodifikation.

Letztlich schließen sich beide Formen nicht aus. Es wäre also durchaus denkbar, dass im Rahmen einer Interventionsmaßnahme sowohl die Verhaltens- als auch die Bedingungsmodifikation genutzt werden, um das gewählte Ziel zu erreichen. Im Rahmen institutionalisierter Bildung und Erziehung, die immer auch unter dem Gesichtspunkt der Effizienz und der Kostenverringerung betrieben wird, ist letztlich auch die Selektionsdiagnostik unumgänglich: Sind die Lernbedingungen wenig veränderbar, die Zuweisung von Förderungsmitteln gering, wird eine möglichst gute Zuweisung zu Bedingungen bzw. eine Auswahl geeigneter Personen zu einer wichtigen Aufgabe der Lehrkraft werden.

Tabelle I (vergl. Wild & Krapp, 2006) gibt einen Überblick über die Struktur der diagnostischen Strategien. Einerseits kann zwischen Selektions- und Modifikationsdiagnostik unterschieden werden, andererseits zwischen dem Objekt der jeweiligen Diagnostik. So kann entweder die Person im Mittelpunkt des diagnostischen Handelns stehen, weil deren Eignung geprüft oder deren Verhalten verändert werden soll. Auf der anderen Seite sind die Umweltbedingungen Gegenstand der Diagnostik, wenn z.B. gefragt wird, welche Bedingungen besonders günstig für den Schüler oder die Schülerin sind oder wie ein Setting didaktisch aufgebaut sein müsste, um beispielsweise eine Schülerin in der Lernentwicklung zu unterstützen.

<table>
<tr><td colspan="4">Diagnostische Strategien</td></tr>
<tr><td colspan="2">Selektionsdiagnostik (häufig institutionell)</td><td colspan="2">Modifikationsdiagnostik (häufig individuell)</td></tr>
<tr><td>Personenselektion</td><td>Bedingungsselektion</td><td>Verhaltensmodifikation</td><td>Bedingungsmodifikation</td></tr>
<tr><td>Objekt der Handlung ist die Person</td><td>Objekt der Handlung ist die Umweltbedingung</td><td>Objekt der Handlung ist die Person</td><td>Objekt der Handlung ist die Umweltbedingung</td></tr>
<tr><td colspan="2">Eher verbunden mit Statusdiagnostik oder Ergebnisdiagnostik</td><td colspan="2">Weist Verbindung zur Prozessdiagnostik auf</td></tr>
</table>

Tabelle I: Diagnostische Strategien

Weiter ist die **Unterscheidung zwischen Status/Ergebnisdiagnostik und Prozessdiagnostik** von Bedeutung. Statusdiagnostik bezeichnet in der Psychologie die Feststellung der Ausprägung eines bestimmten Merkmals (z.B. Intelligenz), das für eine pädagogische oder psychologische Fragestellung relevant ist. In der Pädagogischen Diagnostik wird vorzugsweise von **Ergebnisdiagnostik** gesprochen, weil es hier in der Regel weniger um Eigenschaften geht, sondern um Ergebnisse, die aufgrund des unterrichtlichen Geschehens und des individuellen Lernverhaltens im Sinne eines „Produkts" erfasst werden können. Dies kann zum Beispiel am Ende einer Lerneinheit durch die Überprüfung des Lernerfolgs im Rahmen einer Klausur erreicht werden.

Prozessdiagnostik dagegen hat das Ziel zu prüfen, inwieweit der Lernprozess fortgeschritten ist und wo dieser möglicherweise gezielt unterstützt werden muss, um langfristig den Lernerfolg zu sichern. Hier wäre auch die Überprüfung individueller Fortschritte zum Beispiel im Rahmen eines Verhaltenstrainingsprogramms als Beispiel zu nennen.

Ergebnisdiagnostik ist grundsätzlich eher mit Fragen der Selektion und Zuweisung verknüpft. Zum Beispiel könnte die Beratung einer Schülerin, Physik als Leistungsfach in der Oberstufe anzuwählen, weil ihre Leistungen in den vergangenen Jahren dort überdurchschnittlich waren, sowohl als selektionsdiagnostische Strategie als auch als ergebnisdiagnostisches Vorgehen beschrieben werden. Hingegen die Überprüfung der Wirksamkeit des eigenen Unterrichtsaufbaus durch regelmäßige kleinere Lernzielkontrollen einerseits Teil einer modifikationsdiagnostischen Strategie wäre und dabei andererseits die Prozessdiagnostik nutzt, um Erkenntnisse darüber zu erlangen, was sinnvollerweise modifiziert und angepasst werden sollte, um die Schüler und Schülerin beim Lernen besser zu unterstützen. Insofern könnte man Prozessdiagnostik als kleinschrittige Ergebnisdiagnostik verstehen. Der zentrale Unterschied zwischen beiden liegt jedoch in der Intention: Bei der Ergebnisdiagnostik steht die Leistungsfeststellung im Vordergrund, nicht die Veränderung des Lernprozesses. Die Prozessdiagnostik hingegen hat das Ziel, Lernprozesse zu optimieren, sei es durch zielgerichtete Unterstützung der Lernenden oder durch Veränderung und Anpassung des Unterrichts (also auch der Didaktik).

Ein Problem, dass bei der Prozessdiagnostik auftritt, wird in der Literatur als Häufigkeits-Genauigkeits-Dilemma beschrieben. Damit ist gemeint, dass häufige

Messungen notwendig sind, um einen Prozess abbilden zu können. Gleichzeitig will man aber den Lernprozess selbst nicht ständig stören oder unterbrechen, weshalb in der Regel kurze Tests oder Beobachtungszeiträume gewählt werden. Dies führt dazu, dass die Zuverlässigkeit und auch die Gültigkeit der Messungen nicht immer gewährleistet werden können. Dieses betrifft das Problem der Einhaltung von Gütekriterien, das an anderer Stelle eingehend diskutiert werden soll (S. 30).

DER DIAGNOSTISCHE PROZESS

Stichwörter für dieses Kapitel: diagnostischer Prozess, Aspekte diagnostischen Handelns, Bezugsnormen, Vergleich, Analyse, Interpretation

In diesem Abschnitt beschäftigen wir uns mit der Frage, welche Bereiche zum diagnostischen Handeln gehören und wie der diagnostische Prozess aufgebaut ist. Zunächst lässt sich der diagnostische Prozess in eine bestimmte Abfolge bringen, mit den jeweiligen Stationen sind wiederum unterschiedliche diagnostische Aspekte verbunden.

Folgende Stationen des **diagnostischen Prozesses** lassen sich unterscheiden (Jäger, 2007):

- Fragestellung
- Datenerhebung
- Registrierung
- Interpretation der Daten
- Urteilsbildung
- Urteil

Beispiel für einen diagnostischen Prozess in der Schule

Stellen Sie sich folgende Situation vor: Sie möchten wissen, ob Ihre Schüler und Schülerinnen den Lernstoff verstanden und verinnerlicht haben. Dies wäre zunächst Ihre *Fragestellung*.

Um diese Fragestellung zu beantworten, erheben Sie Daten. Die *Datenerhebung* erfolgt in diesem Fall mit einer mündlichen Prüfung. Während der Prüfung wird ein Protokoll erstellt, das wäre die *Registrierung*. Im Anschluss erfolgt die *Interpretation* der registrierten Daten bezogen auf die Fragestellung (hat der Prüfling den Lernstoff verstanden?). Sie *bilden* sich anhand der registrierten Daten und wahrscheinlich auch anhand Ihrer Erinnerung an die Prüfung ein *Urteil*. Dieses *Urteil* werden Sie anschließend kommunizieren und bekanntgeben, indem Sie nämlich den Prüflingen mitteilen, wie sie abgeschnitten haben.

Am Beginn des Prozesses steht die Fragestellung. Diese kann von unterschiedlicher Komplexität sein. Ist die Frage sehr allgemein, dann kann es notwendig sein, verschiedene Verfahren einzusetzen, um jeweils Teilbereiche der Fragestellung zu beantworten. So wären zum Beispiel bei einer Schülerin, die in manchen Bereichen massive Probleme hat in anderen aber nicht, verschiedene Teilfragen zu untersuchen: Gibt es grundsätzlich Lernprobleme? Gibt es Teilleistungsstörungen? Welche Beziehung hat die Schülerin zu den verschiedenen Lehrkräften? Wie sieht sich die Schülerin selbst? Wie sehen es die Eltern und die Lehrkräfte? Diese Liste von Fragen ist keineswegs vollständig. Unter Umständen führt uns die Beantwortung dieser Fragen noch nicht zur Beantwortung der Ausgangsfrage. Wesentlich ist es also, die Verfahren der Diagnostik auszuwählen, die uns in der Beantwortung der Fragestellung weiterbringen. Dabei kann natürlich auch der Ausschluss einer Diagnose ein Teilergebnis sein. Im genannten Beispiel könnte z.B. die Feststellung, dass keine Teilleistungsstörung vorliegt, ein wichtiges Zwischenergebnis sein.

Kernstück des Diagnostizierens sind also die Erhebung und Auswertung von Daten. Bei der Datengewinnung ist zum Beispiel zu überlegen, welche Verfahren (Tests, Beobachtungen, Gespräche) herangezogen werden können und müssen, um die Fragestellung auch zuverlässig beantworten zu können. Das setzt voraus, dass die Fragestellung prinzipiell mit diagnostischen Mitteln beantwortet werden kann. In unserem Beispiel der mündlichen Prüfung könnte man sich fragen, ob diese tatsächlich geeignet ist, um herauszufinden, ob der Lernstoff verstanden wurde.

Weiter ist sicherzustellen, dass die Registrierung bzw. Datenerhebung objektiv und reliabel erfolgt ist (dazu mehr im Kapitel zu den Gütekriterien einer Messung, S. 30). In unserem Beispiel wäre die Frage zentral, wie das Protokoll der mündlichen Prüfung geführt wurde und ob zum Beispiel auf nicht protokollierte, aber erinnerte Daten zurückgegriffen wurde, wenn zum Beispiel eine Prüferin sagt: „Also bei der Beantwortung der Frage Y erschien mir der Prüfling doch sehr unsicher. Da hat er nur rumgestottert.“ Bei einer Einschätzung nach einer Beobachtung können also etliche Fehler auftreten (Beobachtungsfehler, Erinnerungsfehler etc.).

Bei der Dateninterpretation wiederum wird eingeschätzt, welche Bedeutung das vorgefundene Ergebnis oder Ereignis hat und wie es erklärt werden kann. In der mündlichen Prüfung könnte die Nicht-Beantwortung einer Frage z.B. als

nebensächlich eingeschätzt werden, weil die Prüfer*innen der Ansicht sind, dass der Prüfling die Frage missverstanden hatte, oder aber die Prüfer*innen würden aufgrund von wahrgenommener Prüfungsangst zu der Interpretation kommen, dass ein Prüfling nicht aufgrund mangelnder Vorbereitung, sondern aufgrund seiner Aufregung schlecht abgeschnitten hat. Die Frage, die sich hier stellt, ist, ob und unter welchen Umständen vom Erwartungshorizont abgewichen werden kann und unter Umständen sogar muss, wenn man zu einer fairen und möglichst objektiven Einschätzung gelangen will (vergl. Gütekriterien).

Die Urteilsbildung ist letztlich die Zusammenführung der Erkenntnisse aus der Datengewinnung, Auswertung und Interpretation. Diese ist jedoch nicht unabhängig von der beurteilenden Person. Jäger (2007) stellt dazu fest, dass das Urteil sowohl von der Art und den Quellen der Datengewinnung abhängt, als auch von den Merkmalen des Beurteilenden und den Merkmalen des Urteils selbst.

So unterscheiden sich einzelne Beurteiler*innen in der Informationsverarbeitung (Gedächtniskapazität, Verknüpfung verschiedener Informationsquellen, Gewichtung von Daten) und in ihrer Präferenz für bestimmte Verfahren.

Auch das Urteil selbst kann ganz unterschiedlich gestaltet sein: Sagt es etwas über einzelne Merkmale der Person aus, über die Gesamtpersönlichkeit und wird dieses Urteil mündlich oder schriftlich verfasst? Und schließlich die Frage, welche Funktion mit dem Urteil verbunden ist? Will es nur etwas allgemein beschreiben, vergangenes Verhalten erklären (Retrognose), einen Ist-Zustand beschreiben (Diagnose) oder soll es zukünftiges Verhalten vorhersagen (Prognose)?

Zusammenfassend kann hier also festgehalten werden, dass die Qualität der Urteilsbildung maßgeblich von der **Qualität der Datenerhebung,** auch in Abhängigkeit der **späteren Funktion des Urteils** (Prognose, Zustandsbeschreibung), der Güte der **Verdichtung der Daten** und der **Reflexionsfähigkeit der Diagnostiker*innen** abhängt. Letzteres ist ein kritischer Punkt und zeigt, wie wichtig es ist, als Diagnostiker geschult zu sein und das eigene Handeln immer wieder zu reflektieren. Eine Lehrkraft ist täglich mit diagnostischen Fragen befasst. Die Beantwortung dieser Fragen kann für die Schülerinnen und Schüler zum Teil entscheidend für ihren weiteren Lebens- und Berufsweg sein. Deshalb trägt eine Lehrkraft z.B. bei der Leistungsbeurteilung, aber auch in anderen Fragen der schulischen Diagnostik eine große Verantwortung. Sie muss sich daher wie alle Diagnostiker*innen während des diagnostischen Prozesses der möglichen Urteilsfehler und Verzerrungstendenzen

bewusst sein, um diese zu bearbeiten und möglichst gering zu halten (vergl. Kapitel Gütekriterien).

Aspekte diagnostischen Handelns

Unabhängig vom gewählten Ziel oder vom gewählten Erhebungsinstrument liegen jeder Diagnostik bestimmte Aufgaben zugrunde, die von Diagnostikern gestaltet werden müssen.

Vier Aspekte des diagnostischen Handelns können unterschieden werden (Ingenkamp & Lissmann, 2008):

- Vergleich
- Analyse
- Prognose
- Interpretation

Der Vergleich und die Verwendung von Bezugsnormen

Wann immer wir etwas beobachten oder erhobene Daten auswerten, vergleichen wir das, was wir sehen oder erhoben haben mit anderen Erhebungen, Beobachtungen oder Standards. So könnte man beispielsweise die Anzahl der Fehler in einem Diktat eines Schülers mit der Anzahl seiner Fehler in früheren Diktaten vergleichen. Dies wäre lediglich *eine* Möglichkeit des Vergleichs. Wir könnten aber auch Beobachtungen über eine Person mit Beobachtungen über andere Personen vergleichen, um die gewonnenen Informationen besser einordnen zu können bzw. zu wollen – der Vergleich mit anderen Personen oder Früherem ist ein den Menschen innewohnendes Bedürfnis.

Die Bezugsnormorientierung spielt daher in der Diagnostik, aber vor allem auch bei der schulischen Leistungsbewertung eine wichtige Rolle. Die Bezugsnormorientierung sagt uns, welcher Maßstab bei der Bewertung angesetzt wurde. Bezugsnormen werden in der Regel dann verwendet, wenn eine abschließende Bewertung in Form einer Zensur oder einer Rückmeldung gegeben wird. Während sich aus einer Note nicht herauslesen lässt, welche Bezugsnorm ihr zugrunde lag, lässt sich dies bei verbalen Rückmeldungen häufig eher feststellen. Doch dazu später mehr. Zunächst schauen wir uns die unterschiedlichen Bezugsnormen an.

Grundsätzlich unterscheidet man hierzu im Rahmen Pädagogischer Diagnostik zunächst **drei Bezugsnormen**: Die **sachliche**, die **soziale** und die **individuelle Bezugsnorm** (Rheinberg, 2001). Die inzwischen dazugekommene vierte Bezugsnorm ist die **fähigkeitsbezogene Bezugsnorm** (vergl. Ingenkamp & Lissmann, 2008), die im Rahmen internationaler Vergleichsstudien wie TIMSS und PISA an Bedeutung gewonnen hat.

Sachliche Bezugsnorm

Bei der ***sachlichen Bezugsnorm*** wird das Ergebnis der Datenerhebung an einem zuvor festgesetzten Standard oder einem konkreten Lernziel gemessen.

Ein Beispiel:

Die Prüferin legt vorher fest (z.B. aufgrund von Richtlinien, Curricula und Bildungsstandards), was sie genau erwartet und prüft nun, inwieweit das in der Prüfung erfasste Wissen von diesem gesetzten Lernziel abweicht. Geht es darüber hinaus? Bleibt es deutlich darunter? Wie stark weicht es vom gesetzten Standard (z.B. Curriculum, Lehrplan) ab? Das sind Fragen, die sie sich in diesem Zusammenhang stellen kann. Haben alle das Lernziel erreicht, dann kann man argumentieren, dass die Lehrkraft erfolgreich war. Hier wäre also eine Häufung von Noten im Bereich „sehr gut“ bis „gut“ wünschenswert.

In der sachlichen Bezugsnorm oder auch der kriteriumsorientierten Bezugsnorm werden also bestimmte Standards sowie Grenzen oder Schwellenwerte von Standards inhaltlich begründet. Die erbrachten Leistungen der Schüler und Schülerinnen werden mit einem konstruierten Kriterium verglichen, z.B. mit einer zu erreichenden Kompetenzstufe. Sagt eine Lehrkraft, dass ein Schüler 80 Prozent der Aufgaben gelöst hat, benutzt sie eine sachliche Bezugsnorm. Nutzt eine Lehrkraft ein Kompetenzraster zur Rückmeldung, dann könnte dort zum Beispiel unter der Rubrik „Eigene und fremde Texte bearbeiten können“ folgendes angekreuzt werden: „Ich kann mit Sprache schreibend experimentieren“, weil die Schülerin diese Kompetenz erreicht hat, während vielleicht *kein* Kreuz bei „Ich kann adressatengerecht und sachbezogen schriftlich informieren“ gesetzt wird, da die Schülerin diese Kompetenz noch nicht erreicht hat.

Auch im Rahmen internationaler Schulstudien (vergl. PISA, Baumert et al., 2001; TIMSS, Baumert et al., 1997) wird durch die Nutzung der Kompetenzstufen ähnlich verfahren. Die sogenannte ***fähigkeitsorientierte Norm*** (vergl. Jäger, 2007) als

Spezialform der sachlichen Bezugsnorm beschreibt das Verfahren, die Fähigkeit von Personen in Bezug zu vorgegebenen Kompetenzstufen zu setzen. Letztlich ist dies eine Mischform aus der sachlichen und der sozialen Bezugsnorm: Die gesetzte Norm wird empirisch gewonnen und die Einschätzung der Fähigkeit einzelner wird in Abhängigkeit des Abschneidens aller Personen der Vergleichsgruppe vorgenommen. Durch die Stufung der Kompetenzen wird gleichzeitig sichtbar, welche Stufen noch zu erreichen wären, was wir bereits von der sachlichen Bezugsnorm kennen.

Etwas anders ist es bei der Verwendung von diagnostischen Testverfahren:

Auch ein Intelligenztest oder ein klinischer Test, der darüber entscheidet, ob eine Person z.B. unter ADHS leidet (IQ hier als Ausschlusskriterium im Prozess der Diagnosefindung), vergleicht das erzielte Ergebnis mit einem standardisierten Wert. Dieser Wert entsteht allerdings dadurch, dass auf der Grundlage einer großen Stichprobe (repräsentative Eichstichprobe) zuvor *empirisch* ermittelt wurde, was denn sinnvollerweise erwartet werden kann und was als deutliche Abweichung von einem erwarteten Ergebnis interpretiert werden soll. Strenggenommen vergleichen wir in diesem Fall das Ergebnis der einzelnen Person mit einem Standard, der sich aus der Beobachtung vieler anderer Personen einer repräsentativen Stichprobe zusammensetzt. Der Standard leitet sich aus den Erfahrungen und wissenschaftlichen Erkenntnissen ab, die wir darüber gewonnen haben, was wir als ‚erwartbar# oder auch als ‚normal' interpretieren. Der Normalitätsbegriff an sich ist fragwürdig und wird immer wieder diskutiert (Kemper, 2011). Nichtsdestotrotz kommt ihm bei der Diagnostik eine Bedeutung zu: Erst wenn ein Kind, das z.B. auf ADHS getestet wird, einen bestimmten Wert überschritten hat und damit vom ermittelten Normwert abweicht, wird die Diagnose ADHS gestellt. Gleiches gilt für die Intelligenz: Erst wenn eine Person mindestens zwei Standardabweichungen über dem Mittelwert abschneidet, sprechen wir von einer Hochbegabung bzw. wenn sie zwei Standardabweichungen unter dem Mittelwert liegt, von einer geistigen Behinderung.

Beispiel für eine Bezugsnorm im Rahmen der Testentwicklung

In der Diagnostik wird ein Test entwickelt, der das Merkmal ‚Aufmerksamkeitsdefizit' erfassen soll. Der Test soll bei Kindern und Jugendlichen im Alter von 6 bis 13 Jahre zuverlässig bestimmen, ob ein Aufmerksamkeitsdefizit vorliegt oder nicht. Dazu müssen alters- und geschlechtsspezifische Normen vorliegen. Diese werden dadurch erhoben, dass z.B. eine deutschlandweite Normstichprobe mit ungefähr 800 Kindern und Jugendlichen zwischen 6 und 13 Jahren gezogen wird. Das Ziel ist es, durch eine zufällige Ziehung der Teilnehmer einen realistischen Durchschnittswert und eine realistische Verteilung zu erhalten. Das bedeutet, dass die meisten Teilnehmer*innen unauffällige Werte zeigen werden und nur ein Teil der Stichprobe tatsächlich ein Aufmerksamkeitsdefizit aufweisen wird. Anhand der Verteilung der Werte erhält der Testentwickler Auskunft darüber, wie das Merkmal „normalerweise" verteilt ist, was in unterschiedlichen *Altersklassen* und bezogen auf das *Geschlecht* an Aufmerksamkeitsleistung erwartet werden kann. Da die meisten psychischen Merkmale normalverteilt[2] sind, entsteht eine Glockenkurve: Das bedeutet, dass das Merkmal symmetrisch um den Mittelwert verteilt ist und dass nur ein kleiner Teil der Stichprobe extreme Ausprägungen des Wertes aufweist (sowohl in die Richtung hohe Aufmerksamkeit als auch in die Richtung Aufmerksamkeitsdefizit).

Die Testentwickler werden einen Cut-off Wert festlegen, der darüber entscheidet, ab welchem Wert eine Person, die den Test absolviert, im Folgenden als Person mit einem Aufmerksamkeitsdefizit bestimmt wird. Üblicherweise legt man den Cut-off-Wert so fest, dass einer Person ein Aufmerksamkeitsdefizit zugesprochen wird, wenn sie einen Wert aufweist, der höher ist als der Wert von 95% der Stichprobe. Die Bewertung orientiert sich hier also einerseits an einer sozialen Vergleichsgruppe (der Normstichprobe), aber sie orientiert sich auch an einem **Kriterium**: Die Testentwickler werden bei der Testentwicklung sorgfältig überlegt und theoretisch und empirisch begründet haben, warum bestimmte Anzeichen von fehlender Aufmerksamkeit als Aufmerksamkeitsdefizit betrachtet werden und andere Anzeichen dagegen als „normal" gewichtet werden.

[2] „Die Normalverteilung (oder Gauß'sche Glockenkurve) ist eine bei biologischen, psychologischen und soziologischen Variablen häufig zu beobachtende Idealform einer Häufigkeitsverteilung. Sie ist dadurch gekennzeichnet, dass mittlere Ausprägungen einer Variable am häufigsten vorkommen, während extreme Merkmalsausprägungen sehr selten sind. Grafisch dargestellte Normalverteilungen sind symmetrisch und haben einen glockenförmigen Verlauf. Eine Normalverteilung ist dann zu erwarten, wenn eine Variable von zahlreichen Faktoren beeinflusst wird, die voneinander unabhängig sind und additiv zusammenwirken. Dies gilt beispielsweise für Schulleistungen oder Intelligenz" (Preiser, 2003, S.57).

Soziale Bezugsnorm

Von ***sozialer Bezugsnorm*** sprechen wir, wenn ein Ergebnis mit den Ergebnissen anderer aus der gleichen sozialen Gruppe (z.B. der gleichen Klasse) verglichen wird. Das Urteil, ob jemand also besonders gut oder schlecht in einem Test abgeschnitten hat, hängt hier davon ab, wie die Gruppe insgesamt abgeschnitten hat. Die soziale Bezugsnorm wird häufig in der Schule verwendet, wenngleich sie große Probleme birgt: Wenn das Urteil über die Fähigkeit eines Einzelnen in Abhängigkeit der Leistungsfähigkeit der Gruppe gemessen wird, so kann die individuelle Leistungsfähigkeit systematisch unterschätzt oder überschätzt werden (Big-Fish-Little-Pond-Effect).

Ist die Gruppe nämlich besonders stark, dann muss der Einzelne mehr Leistung erbringen, um den Durchschnitt zu erreichen als dies der Fall ist, wenn die Gruppe durchschnittlich oder sogar leistungsschwach ist. Eine Person kann dann vor dem Hintergrund verschiedener Gruppenleistungen mit *derselben* Leistung als leistungsschwach, durchschnittlich oder leistungsstark angesehen werden. Zudem ist die soziale Bezugsnorm aus motivationalen Gründen ungünstig: Gerade leistungsschwächere Schüler und Schülerinnen werden Probleme haben ihre Fortschritte zu erkennen, wenn sie gemessen an einer leistungsstarken Gruppe unabhängig von ihrem Fortkommen immer wieder am unteren Ende der Bewertungsskala stehen, weil die anderen eben „besser“ sind.

Beispiel

Klasse A hat im Durchschnitt 60 % der Aufgaben gelöst, Klasse B hat im Durchschnitt 75 % der Aufgaben gelöst. Der Schüler aus Klasse A bekommt mit 75% richtiger Aufgaben eine „2“. Der Schüler aus Klasse B bekommt bei einer Bewertung auf Grundlage der sozialen Bezugsnorm mit 75% richtiger Aufgaben eine „3“. In Abhängigkeit von der durchschnittlichen Leistung der Bezugsgruppe wird eine Person mit gleicher Leistung also besser oder schlechter bewertet.

Individuelle Bezugsnorm

Den oben genannten Kritikpunkt greift die ***individuelle Bezugsnorm*** auf. Bei der individuellen Bezugsnorm wird das Ergebnis des Einzelnen mit seinen vorher erzielten Ergebnissen verglichen. Man schaut auf die Entwicklung des Einzelnen. Es werden also Leistungssteigerungen und Leistungsabfälle betrachtet. Aussagen wie „Du hast dich in letzter Zeit viel häufiger gemeldet als vorher“ oder „Es gelingt dir immer besser in Konfliktsituationen ruhig zu bleiben.“ wären Beispiele für eine individuelle Bezugsnorm.

Das Verfahren ist insofern aufwändiger als die Nutzung der sozialen Bezugsnorm, als dass die Lehrkraft individuell Entwicklungen betrachten und dokumentieren muss.

Dies ist in der Diagnostik unumgänglich, wenn man z.B. einen Förderbedarf festgestellt hat und im Anschluss an eine durchgeführte Intervention prüfen möchte, ob sich diese positiv auf die Entwicklung der Person ausgewirkt hat. Aber auch im Rahmen der Leistungsprüfung in der Schule hat diese ihre Berechtigung: Schüler und Schülerinnen können aufgrund der Verwendung der individuellen Bezugsnorm eigene Lernfortschritte wahrnehmen. Auch hinsichtlich der *Aufgabenstellung* kann eine Anpassung des Schwierigkeitsgrades an den Entwicklungsstand des Schülers oder der Schülerin sinnvoll sein. Nachweislich unterstützt der Einsatz individueller Bezugsnorm den Aufbau eines günstigen Fähigkeitsselbstkonzepts, weil die Misserfolgsmotiviertheit abnimmt und variable (also beeinflussbare) Kausalattributionen zur Erklärung des eigenen Erfolgs oder Misserfolgs herangezogen werden (Dickhäuser & Rheinberg, 2003).

Ein Beispiel

Eine Schülerin hält ein Referat im Physikunterricht. Sie nehmen wahr, dass sie anders als bei den letzten Malen deutlich freier spricht. Sie geben ihr ein positives Feedback und fragen die Schülerin, ob sie sich diesmal anders als bei den letzten Malen auf das Referat vorbereitet hat. Die Schülerin erzählt Ihnen, dass sie mehrfach geübt hat, ihrem kleinen Bruder und ihrer Freundin physikalische Zusammenhänge in eigenen Worten zu erklären. Sie loben dies ausdrücklich und stellen einen Zusammenhang her zwischen der Fähigkeit frei zu sprechen und dem Üben des freien Erklärens. Dadurch verstärken Sie die Schülerin in der Verknüpfung des von ihr gezeigten, kontrollierbaren Verhaltens mit dem Erfolg („frei sprechen können“). Sie entwickelt dadurch eine günstige Ursachenzuschreibung (Kausalattribution), weil sie diese Ursache selbst beeinflussen und steuern kann.

Vor- und Nachteile von Bezugsnormen

Die Bevorzugung bestimmter Bezugsnormen geht in der Regel mit spezifischen Ursachenerklärungen einher: Bei der sozialen Bezugsnorm führt die Lehrkraft den Erfolg der Einzelnen häufig auf deren Fähigkeiten und Begabungen zurück. Die soziale Bezugsnorm wird daher auch zur Selektion herangezogen (Bestenauswahl, Bewerberauswahl). Bei der individuellen Bezugsnorm findet eine Anpassung der Aufgaben an die jeweilige Lernausgangslage statt und es wird angenommen, dass durch Anstrengung Lernfortschritte erzielt werden können.

Sowohl die sachliche als auch die soziale Bezugsnorm können keine Lernfortschritte dokumentieren. Sie sind daher gerade für schwächere Schüler und Schülerinnen häufig demotivierend. Letzteres gilt in besonderem Maße für die soziale Bezugsnorm: gerade in durchschnittlich leistungsstarken Klassen werden schwächere Schüler und Schülerinnen bei Anwendung der sozialen Bezugsnorm besonders demotiviert, da sie auch bei Lernfortschritten in der Regel nicht ihren Rang innerhalb der Bezugsgruppe verändern und so weiterhin schlecht bewertet werden, wenn ihre Leistung unterhalb des Durschnittes der (leistungsstarken) Klasse bleibt. Die individuelle Bezugsnorm ist hier deutlich motivierender, da sie Fortschritte anerkennt. Bei alleiniger Anwendung blendet sie allerdings nicht nur Leistungsunterschiede, sondern unter Umständen auch das Verfehlen von Kompetenzzielen aus (Rheinberg, 1980).

Kritiker der individuellen Bezugsnorm argumentieren zudem, dass die Verwendung der individuellen Bezugsnorm dazu führt, dass Schüler und Schülerinnen sich selbst fehleinschätzen, weil sie ihre Leistung systematisch überbewerten. Ihr Argument ist, dass z.B. ein Schüler, der im selben Diktat erst vierzig Fehler, später aber dreißig Fehler macht, nach wie vor schlecht abschneidet, auch wenn er sich verbessert hat. Außerdem wäre die alleinige Nutzung der individuellen Bezugsnorm ungerecht, da Schüler und Schülerinnen, die von Beginn an sehr gute Leistungen zeigen, sich ja nicht in dem Maße verbessern können wie leistungsschwächere Schüler und Schülerinnen.

Eine Lösung für dieses Problem bietet die Nutzung sachlicher Kriterien und Orientierung an Kompetenzstufen bei gleichzeitiger Rückmeldung individueller Fortschritte: Dies stellt letztlich die für Schüler und Schülerinnen sinnvollste Bezugsnormorientierung dar.

Einerseits können Lernende erkennen, ob und wie weit sie von einem gesetzten Ziel entfernt sind, andererseits nehmen sie ihre Lernfortschritte aber auch ihre Leistungsabfälle wahr. Lehrkräfte profitieren von der Verwendung dieser Bezugsnormen ebenfalls, weil sie ihren Unterricht am Erreichen der vorgegebenen Kompetenzziele ausrichten können. Gleichzeitig können sie differenziert auf die unterschiedlichen Lernvoraussetzungen und Lernstände der Schüler und Schülerinnen eingehen.

Die soziale Bezugsnorm ist zumindest im schulischen Bereich auf der Ebene des Vergleichs mit konkreten Mitschülern und Mitschülerinnen und auf Klassenebenen verzichtbar und eher von Nachteil. Sie zeigt ihre Berechtigung vorwiegend im diagnostischen Rahmen, wenn es darum geht, deutliche Abweichungen von erwarteten Normwerten (z.B. bei Sprachstandserhebungen etc.) zu diagnostizieren und damit einen Förderbedarf festzustellen (s. Beispiel unten).

Die Analyse

Nach einer vergleichenden Einordnung unserer Beobachtungen, Daten oder Testergebnisse, stellt sich die Frage, *wodurch* das festgestellte Verhalten oder Ergebnis erklärt werden kann.

Weicht ein Ergebnis zum Beispiel deutlich vom erwarteten Standard ab, dann fragen wir uns, warum das so ist. Die Analyse ist das Kernstück des diagnostischen Prozesses, da hier Hypothesen darüber aufgestellt werden, warum etwas besonders gut oder auch gar nicht gelingt. Entsprechend unserer Schlüsse, die wir aus den Beobachtungen und Daten ziehen, werden wir (Förder-)Maßnahmen einleiten, die, wenn unsere Analyse korrekt war, greifen können oder aber keinen Erfolg erzielen werden, wenn unsere Analyse fehlgegangen ist.

Warum schneidet ein Prüfling gut oder schlecht ab? Diese Frage zu beantworten, ist in den wenigsten Fällen trivial, da die Anzahl der Einflussfaktoren auf die Leistung des Prüflings groß ist.

So könnten Gründe für das Abschneiden in der individuellen Prüfungsvorbereitung liegen. Häufig begründen wir Misserfolg genau damit, dass wir annehmen, dass sich die Prüflinge nicht angemessen oder intensiv genug vorbereitet haben. Wir könnten uns aber auch fragen, ob die Materialien, die wir zur Vorbereitung gegeben haben, sinnvoll und ausreichend waren, oder ob wir transparent gemacht haben, was wir in der Prüfung erwarten.

Zudem stellt sich die Frage nach den individuellen Voraussetzungen der Schüler und Schülerinnen. Das könnten einerseits Persönlichkeitsmerkmale (Intelligenz, Konzentrationsfähigkeit etc.) sein, aber auch die häuslichen Umgebungsbedingungen (Zugang zu Informationen, Medien etc.; Unterstützung durch Eltern, Nachhilfe, vergl. Bourdieu, 1002).

Andererseits könnte ich als Lehrkraft aber auch überlegen, ob mein Unterricht zu dem Prüfungsergebnis beigetragen hat: Vielleicht habe ich einen bestimmten Themenbereich didaktisch schlecht aufbereitet oder zu wenig Zeit für die Einübung von notwendigen Routinen verwendet. War er adressatenorientiert und hat die Bedürfnisse und Vorkenntnisse der Schüler und Schülerinnen berücksichtigt?

Ein dritter Bereich wäre die Frage nach der Prüfung selbst: Wie sind die Rahmenbedingungen? Liegt die Prüfung in der sechsten Stunde? Gab es Störungen oder habe ich deutlich Prüfungsangst wahrgenommen? Habe ich durch mein Verhalten das Prüfungsergebnis beeinflusst? Waren die Prüfungsfragen verständlich? Habe ich Rückfragen zugelassen oder nicht?

Aus dieser Aufzählung, die keinesfalls vollständig ist, wird deutlich, wie komplex die Analyse sein kann. Dennoch ist es sinnvoll, sich grundsätzlich gerade die Fragen

zu stellen, die die Lehrkraft und ihren Unterricht selbst betreffen: Fragen, inwieweit die Prüfung angemessen gestaltet war, wie Vorbereitung und Unterricht zu den erzielten Ergebnissen beigetragen haben. Das gilt auch für die Analyse besonders gelungener Bedingungen. Warum haben meine Schüler und Schülerinnen beim Thema X besonders gut abgeschnitten? Das zu klären, kann dazu beitragen, den Unterricht und den Lernprozess langfristig immer wieder zu verbessern.

Die Analyse individueller Bedingungen dagegen wird in der Regel dann vorgenommen, wenn Probleme entstehen oder sich häufen und eine Intervention angeraten erscheint. Dann ist umfassender zu klären, was zu dem beobachteten Verhalten oder Testergebnis geführt hat.

Die Prognose

Mit der Prognose wird eine Vorhersage darüber getroffen, welches Verhalten in der Zukunft, ausgehend vom jetzigen Verhalten, zu erwarten ist. Wenn eine Lehrkraft darüber urteilt, ob ein Schüler zum Abitur zugelassen werden kann, dann prognostiziert sie dessen Erfolg auf Grundlage seiner bisherigen Leistungen.

Im schulischen Alltag stellen Lehrkräfte häufig auch intuitiv Prognosen, z.B. in Form von Annahmen. Sie nehmen beispielsweise im Englischunterricht an, dass sie sich lange und intensiv genug mit dem Thema Konditionalsätze auseinandergesetzt haben und deshalb die Schüler und Schülerinnen in der Lage sein müssten, dieses zu beherrschen. Oder Sie gehen aufgrund einer Mobbingpräventionsveranstaltung davon aus, dass die Schüler und Schülerinnen für das Thema sensibilisiert worden sind, und Übergriffe zukünftig melden.

Inwiefern Prognosen zutreffend oder angemessen sind, hängt von der Qualität der Analyse und der erhobenen Daten ab. Festzustellen ist jedoch, dass die Lehrkraft sich bestimmten gesellschaftlichen Erwartungen ausgesetzt sieht, die sie dazu zwingen *Beurteilungen* vorzunehmen, die immer auch prognostischen Charakter haben.

Die Interpretation

Nach Vergleich, Analyse und Prognose widmen wir uns nun der Interpretation der Daten, der abschließenden Beurteilung. In diese Beurteilung fließen in der Regel nicht nur die Erkenntnisse und Informationen aus z.B. einer Klassenarbeit, sondern auch Einstellungen und Erwartungen sowie frühere Erfahrungen ein. Weiter können eine oder mehrere Quellen genutzt werden. So könnten bei der Interpretation der Prüfungsleistung in einer mündlichen Prüfung auch Informationen aus den im Rahmen des Unterrichts geleisteten Beiträge des Prüflings eine Rolle spielen. Zudem ist es entscheidend, ob die Lehrkraft sich ausschließlich auf die Interpretation der eigenen Wahrnehmung stützt oder aber auch Fremdbeobachtungen und objektive Verfahren nutzt. In beiden Fällen können Fehler auftreten. So fällt bei ausschließlicher Nutzung der eigenen Wahrnehmungen das Korrektiv einer Fremdbeurteilung oder anderer Perspektiven weg, zudem fallen Datensammlung und Wertung zusammen (vergl. Ingenkamp & Lissmann, 2008), was methodisch fragwürdig ist. Im zweiten Fall ist genau zu prüfen, welche Güte die hinzugezogenen Quellen haben. Hier besteht die Gefahr, dass Fremdurteile übernommen werden, ohne dass sie hinsichtlich ihrer Qualität geprüft worden sind.

Der Interpretationsvorgang ist also mit vielen Fehlern behaftet. Schrader und Helmke (2001, S. 8) schreiben dazu: „[….], dass Lehrerurteile häufig nicht den Anforderungen genügen, die man an professionelle Diagnosen stellen muss (zusammenfassend Tent, 1998)“.

Untersuchungen zeigen aber auch, dass Lehrkräfte die Rangreihe in ihrer Klasse vergleichsweise zutreffend einschätzen. Gleichzeit allerdings muss man feststellen, dass gleichen Noten in unterschiedlichen Klassen, ganz unterschiedliche Leistungen zugrunde liegen (Ingenkamp, 1997), was wiederum hochproblematisch ist.

Trotz aller Kritik und Vorbehalte ist die Beurteilung von Schülerinnen und Schülern eine Aufgabe, deren Gestaltung von der Lehrkraft aus unterschiedlichen Gründen erwartet wird. Bei der Beurteilung einer Person lassen sich verschiedene Begründungszusammenhänge unterscheiden (Jäger, 2007):

Gesellschaftliche Implikationen beschreiben die Funktionen, die einer abschließenden Interpretation und daran anschließenden Prognose im Rahmen einer Beurteilung zugewiesen werden. So soll sie die Verteilung von (begrenzten) Qualifizierungsmöglichkeiten sinnvoll regeln. Lernende sollen gemäß ihren

Fähigkeiten und Begabungen Zugang zu Qualifizierungsmöglichkeiten erlangen unter gleichzeitiger Berücksichtigung des gesellschaftlichen Bedarfs.

Didaktische Implikationen beschreiben das Ziel, durch Beurteilungen und Prognosen Lehr-Lern-Prozesse zu optimieren und zu steuern, Lernende zu beraten und Lehr- und Lernprozesse hinsichtlich ihrer Qualität zu bewerten.

Als persönliche Implikationen einer Beurteilung können einerseits die Rückmeldungsfunktion, andererseits aber auch die damit einhergehende personale und soziale Anerkennungsfunktion beschrieben werden. Die Beurteilten bekommen im Rahmen der Prognose und der damit einhergehenden Beurteilung mitgeteilt, wie sie gesehen und bewertet werden und was ihnen zugetraut wird. Dies kann anerkennend, aber auch kränkend erlebt werden.

Abschließend lässt sich also feststellen, dass der diagnostische Prozess von hoher Komplexität bei gleichzeitiger Fehleranfälligkeit gekennzeichnet ist. Ihn sinnvoll zu gestalten ist eine der zentralen Aufgaben einer Lehrkraft. Wie kann dieser Prozess optimiert werden? Welche Beurteilungsfehler gibt es? Wie können diese Fehler vermieden werden? Mit diesen Fragen beschäftigt sich das folgende Kapitel.

GÜTEKRITERIEN

Wenn Sie ein Thermometer benutzen, um sich einen Eindruck davon zu machen, was eine angemessene Kleidung ist, um vor die Tür zu gehen, dann möchten Sie sich auf die Anzeige verlassen können. Würde das Thermometer beispielsweise bei hoher Luftfeuchtigkeit immer ein paar Grad mehr anzeigen, dann würden Sie es als *unzuverlässig* einstufen. Die Messung wäre also fehlerbehaftet und für Sie von geringem Wert.

Auch im Rahmen der Pädagogischen Diagnostik führen wir Messungen durch, die in der Regel mit größeren oder kleineren Messfehlern behaftet sind. Als Gütekriterien bezeichnet man Maße, an denen sich die Qualität des diagnostischen Verfahrens von der Erhebung bis zur Beurteilung bemisst.

GÜTEKRITERIEN

Hauptgütekriterien			Nebengütekriterien
Objektivität	**Reliabilität**	**Validität**	
Durchführung	*Interne Konsistenz*	*Inhaltsvalidität (Unterrichtsvalidität und Curriculare Validität)*	*Normierung*
Auswertung	*Stabilität*	*Konstruktvalidität*	*Vergleichbarkeit*
Interpretation		*Kriteriumsvalidität (Übereinstimmungsvalidität, Vorhersagevalidität)*	*Ökonomie*
			Nützlichkeit
			Fairness

Nebengütekriterien

Man unterscheidet Haupt- und Nebengütekriterien. Nebengütekriterien werfen Fragen vor allem bezogen auf standardisierte Tests auf, z.B. inwieweit die Einführung eines neuen Testverfahrens sinnvoll ist.

Ist es ökonomischer als ein altes Verfahren, weil es günstiger oder einfacher in der Handhabung ist? Diese Frage betrifft das Nebengütekriterium *Ökonomie*. Es wäre zum Beispiel dann erfüllt, wenn ein altes Verfahren sehr aufwändig ist, ein neueres, inhaltlich gleichwertiges Verfahren dagegen mit deutlich geringerer Testzeit und geringerem Auswertungsaufwand zum gleichen Ergebnis kommt.

Kann man das Testverfahren mit anderen Verfahren gleichsetzen? Diese Frage thematisiert die *Vergleichbarkeit* von Verfahren. Also die Frage, ob z.B. zwei Tests zur Erfassung von Lese-Rechtschreibschwächen das gleiche Konstrukt in vergleichbarer Qualität erfassen können.

Ist das Verfahren normiert worden? Diese Frage bezieht sich auf die Erstellung von Normtabellen und gibt Aufschluss darüber, ob die Eichstichprobe, anhand derer die Standardisierung erfolgte, tatsächlich repräsentativ bezogen auf die Zielgruppe ist (dazu siehe das Beispiel unten).

Beispiel zum Thema Normierung

In der Geschichte der Psychologie gab es immer wieder Fälle, wo bei der Eichung eines Instruments eine nicht repräsentative Stichprobe verwendet wurde. So wurde in den fünfziger Jahren ein Persönlichkeitstest entwickelt, dessen Standardisierung, also die Zuweisung der Werte zu den Bereichen ‚normales' versus ‚krankhaftes' Verhalten anhand einer Stichprobe vorgenommen wurde, die nur aus Patienten einer Nervenheilanstalt und ihrer Pfleger (!) vorgenommen wurde. Die Frage, ob die Pflegekräfte tatsächlich in ihrer Persönlichkeitsstruktur repräsentative Vertreter der Grundgesamtheit der ‚nicht-Erkrankten' waren, darf mit Recht gestellt werden.

Auch im Zuge der Entwicklung von Intelligenztests kam es immer wieder zu Problemen der Normierung, weil in den fünfziger Jahren beispielsweise viele Tests an weißen Männern normiert wurden, Frauen und Vertreter anderer Ethnien hingegen nicht Teil der Eichstichprobe waren. Dies führt in beiden Fällen zu einer fragwürdigen Normierung, weil die Standardsetzung eben anhand einer spezifischen Untergruppe erfolgte und deren Mittelwerte als „Norm" für alle gesetzt wurden.

Ist das Verfahren nützlich, weil es beispielsweise uns hilft, schnell und sicher Probleme zu diagnostizieren? Die Frage der *Nützlichkeit* stellt sich zum Beispiel bei Screeningtests, die uns helfen sollen, frühzeitig Entwicklungsprobleme zu diagnostizieren.

Und abschließend die Frage, ob das Testverfahren fair ist? Das Kriterium der Fairness ist auch im Rahmen von *Pädagogischer* Diagnostik von großer Bedeutung. Seine Einhaltung bedeutet, dass prinzipiell alle Teilnehmer die gleichen Chancen haben, im Test gut abzuschneiden. Im Rahmen von Intelligenztests beispielsweise versuchen sogenannte „Culture-Free-Tests" möglichst auf sprachliche Aufgaben zu verzichten, weil diese Menschen aus bildungsnahen Schichten bevorzugen. Auch die thematische Wahl von bestimmten Themenbereichen kann dazu führen, dass zum Beispiel Mädchen oder Jungen bevorzugt werden. Bei der Entwicklung von fairen Tests wird also darauf geachtet, dass die Ergebnisse nicht durch Bildung, soziale Herkunft, Geschlecht oder Religionszugehörigkeit beeinflusst werden. Sind zum Beispiel PISA-Aufgaben unabhängig vom Fach komplex formuliert, zeigt sich,

dass die Testergebnisse über die verschiedenen Fächer stark miteinander korrelieren. Hier sehen wir zugleich eine Verletzung eines weiteren Gütekriteriums – der Validität (s. unten), weil der Test dann weniger fachliche Inhalte erfasst, als vielmehr die Lesekompetenz. Doch dazu später mehr.

Hauptgütekriterien

Die Hauptgütekriterien sind für jegliche Diagnostik und daher sowohl für die Leistungsdiagnostik als auch für die Pädagogische Diagnostik bedeutsam. Auch Lehrkräfte sollten versuchen, diese möglichst einzuhalten. Warum dies bedeutsam ist, lässt sich unter anderem daran zeigen, dass die Verletzung des Gütekriteriums Objektivität nachfolgend automatisch dazu führt, dass die Einhaltung der anderen Gütekriterien eingeschränkt ist.

Beispiel

Wenn Sie einen Test unkonzentriert auswerten und dabei grobe Fehler machen, wäre dies zunächst eine Verletzung des Gütekriteriums Auswertungsobjektivität. Sie vergeben in Folge der Fehler falsche Punktzahlen und schließen daraus fälschlich auf z.B. Können oder fehlendes Können (Interpretationsobjektivität). Dieses Ergebnis wird zudem wenig zuverlässig (reliabel, vergl. Reliabilität) sein, denn Sie haben ja falsch ausgewertet. Oder anders gesprochen: Sollten Sie den Test wiederholen und ihn diesmal korrekt auswerten, werden die Schüler bei gleicher Beantwortung *andere* Ergebnisse erzielen. Auch die Validität ist eingeschränkt, denn Sie haben ja gerade durch die Fehler nicht messen können, wie gut die Schüler und Schülerinnen den Unterrichtsgegenstand verstanden haben. Allenfalls haben Sie (zumindest theoretisch) erfasst, wie viele Fehler *Sie selbst* durchschnittlich machen, wenn Sie einen Test unter der Bedingung „mangelnde Konzentration“ auswerten. Sie sehen also, eine Verletzung des einen Gütekriteriums Objektivität schlägt durch auf die anderen nachfolgenden Gütekriterien der Reliabilität und Validität.

Eine Verletzung der Reliabilität führt automatisch wiederum zu einer eingeschränkten Validität. Wenn Sie nämlich zum Beispiel in einem Vokabeltest nur den ersten Block (6 Vokabeln) von insgesamt 4 Blöcken (30 Vokabeln) zu lernende Vokabeln abfragen, dann werden Sie nicht zuverlässig erfasst haben

können, ob die Schüler und Schülerinnen alle Blöcke gelernt haben. Das könnte sich auch auf die Validität durchschlagen. Es könnte nämlich sein, dass Sie auf diese Weise vor allem das *Arbeitsverhalten* der Schüler und Schülerinnen erfassen: Schüler*innen, die mit dem ersten Block angefangen haben und diesen gründlich gelernt haben, sonst aber noch nichts, würden besser abschneiden als Schüler und Schülerinnen, die versucht haben sich alle Vokabeln auf einmal einzuprägen.

Andersherum gilt dies nicht: Nehmen wir an, Sie stellen sich auf eine Waage, die zuverlässig Gewicht misst und lesen Ihren Wert ab. Das wiederholen Sie alle zwei Tage und notieren sich das Gewicht. Sie erfassen das Gewicht objektiv (Sie schummeln nicht und Sie messen immer zur gleichen Tageszeit, direkt nach dem Aufstehen ohne Bekleidung, ihre Freundin liest die gleichen Werte ab wie Sie) und reliabel (weil die Waage geeicht ist), aber nach acht Wochen schauen Sie sich Ihre Gewichtskurve an und sagen: „Die relativ konstanten Werte zeigen, dass ich gesund bin!“ Hier wäre die Validität verletzt. Denn konstantes Körpergewicht ist kein hinreichender Beleg für „Gesundheit“. Das Instrument taugt also nicht dazu, zu prüfen, ob Sie wirklich gesund sind. Dennoch – und das ist jetzt anders als im Beispiel davor – sind die Gütekriterien Objektivität und Reliabilität nicht verletzt. Ihre Notizen stimmen ja, nur messen Sie nicht das, was Sie denken, dass sie messen. Wir befassen uns im Folgenden mit den einzelnen Gütekriterien und sehen uns diese genauer an.

Wichtig ist jedoch sich zu merken, dass ein Test oder ein diagnostisches Verfahren niemals gültig oder valide sein kann, wenn die Reliabilität oder die Objektivität deutlich eingeschränkt sind. Weiter kann auch die Reliabilität nicht hoch sein, wenn die Objektivität bereits deutlich eingeschränkt ist.

Objektivität

Zunächst beginnen wir mit dem Gütekriterium der Objektivität. Als objektiv bezeichnet man eine Messung immer dann, wenn die Messergebnisse möglichst unabhängig vom Untersucher sind, bzw. wenn unabhängige Untersucher*innen zum gleichen Ergebnis kommen, wenn sie dasselbe Merkmal erheben.

Es gibt drei unterschiedliche Formen der Objektivität: Durchführungsobjektivität, Auswertungsobjektivität und Interpretationsobjektivität.

Durchführungsobjektivität beschreibt, inwieweit für alle Teilnehmenden an einer Prüfungs- oder Testsituation die gleichen Bedingungen herrschen. Wenn wir uns

also darum bemühen, während einer Klausur alle möglichen Fragen vorab zu klären, wir allen ein möglichst ungestörtes Arbeiten garantieren, dann wären das Beispiele dafür, dass wir uns um Durchführungsobjektivität bemühen. Dabei ist zu bedenken, dass wir niemals alle Faktoren, die die Durchführungsobjektivität betreffen, auch kontrollieren können:

So können Sie beispielsweise vielleicht nicht beeinflussen, dass die Arbeit in der fünften Stunde nach dem Sportunterricht geschrieben wird, während die andere Klasse in der ersten Stunde schreibt. Auch können unterschiedliche Zeiten je nach Schüler günstiger oder ungünstiger sein. Während manche in der ersten Stunde ein Leistungshoch haben, sind andere vielleicht noch nicht richtig wach. Wen setzen Sie neben wen? Vielleicht setzt eine Lehrkraft vor der Klausur einige männliche Schüler auseinander, die sie verdächtigt, womöglich abzuschreiben, während andere nebeneinander sitzen bleiben dürfen. Auch hier ist die Durchführungsobjektivität beeinträchtigt, da schon allein das unterschiedliche Verhalten der Lehrkraft gegenüber den Schülern und Schülerinnen dazu führt, dass manche vielleicht verunsichert oder verärgert in die Prüfung gehen. Es geht also hier darum, die Prüfungsbedingungen für alle Beteiligten so gut es geht vergleichbar zu gestalten.

Die *Auswertungsobjektivität* befasst sich mit der Frage, ob unterschiedliche Auswerter*innen dieselbe Arbeit/Beobachtung oder ein gezeigtes Verhalten gleich interpretieren. Wenn Sie zum Beispiel einen Multiple-Choice-Test verwenden, ist die Wahrscheinlichkeit für eine hohe Auswertungsobjektivität normalerweise groß, denn die Auswerter*innen müssen nur entscheiden, ob das Kreuz an der richtigen Stelle gemacht wurde.

In dem Moment, wo Sie aber zum Beispiel einen Aufsatz hinsichtlich seines Aufbaus und seiner sprachlichen Qualität beurteilen, werden mit hoher Wahrscheinlichkeit unterschiedliche Auswerter*innen zu unterschiedlichen Einschätzungen gelangen. Ähnliches gilt für mündliche Prüfungen. Vielleicht interpretiert ein Prüfer das schnelle Sprechen des Prüflings als Beleg für sein Können. Eine andere Prüferin dagegen sieht dies als Ausdruck für auswendiggelerntes, heruntergebetetes Wissen ohne tiefere Reflexion an. In diesem Zusammenhang spielen auch sogenannte Beurteilungsfehler (vergl. ab S. 44) eine große Rolle.

Die *Interpretationsobjektivität* schließlich zeigt, inwieweit mehrere Beurteiler*innen das *gleiche Ergebnis gleich* interpretieren. Gesetzt den Fall also, dass wir uns alle einig sind, dass eine Schülerin in einem Essay im Fach Geschichte bestimmte historische Fakten und Ereignisse korrekt dargestellt hat, dass aber andererseits sinnvolle Querverweise und Interpretationen fehlen, dann würde bei einer eingeschränkten Interpretationsobjektivität eine Prüferin vielleicht dieses Ergebnis als „ausreichend" bezeichnen, da aus ihrer Sicht nur fundamentale Wissensbestandteile wiedergegeben wurden, aber eine Eigenleistung fehlt. Ein zweiter Prüfer dagegen würde zu der Ansicht gelangen, dass die Schülerin das Thema verstanden und sinnvoll bearbeitet hat, wenngleich einige weiterführende Gedanken schön gewesen wären und beurteilt die Leistung als „gut". Einerseits gibt es hier den sogenannten *pädagogischen Ermessensspielraum*: Es mag nämlich gute Gründe geben, warum wir uns entscheiden ein Gesamtwerk positiver wahrzunehmen als es die Summe der Einzelteile vermuten ließe.

Beispiel

Jonathan hat einen Aufsatz geschrieben. Ihnen fällt auf, dass er, obwohl Sie die wörtliche Rede geübt haben, auf Einschübe („sagte er", „meinte sie", „erwiderte Tim" etc.) verzichtet hat. Dies führt dazu, dass sein Dialog im Aufsatz eine gewisse Dynamik bekommt und den Charakter (Schlagabtausch) des Dialogs sinnvoll unterstützt.

Laut Ihrem Erwartungshorizont müssten Sie Jonathan Punkte abziehen, weil die eingeübte Regel verletzt worden ist. Sie erkennen aber, dass diese Regelverletzung als persönliches Stilmittel eingesetzt wurde und bewerten diese positiv.

In dem genannten Beispiel führt der pädagogische Ermessensspielraum zu einer Missachtung des zuvor formulierten Erwartungshorizonts. Dieser Ermessensspielraum wird aber nicht willkürlich genutzt, sondern begründet wahrgenommen. Das ist letztlich entscheidend und kann uns davor bewahren, Fehler zu begehen, die aufgrund einer eingeengten Wahrnehmung (vergl. Kapitel zum Thema Beobachtungs- und Beurteilungsfehler) oder bestimmter rigider Erwartungen entstehen könnten.

Auch die Frage, inwieweit wir von konkret vergebenen Punkten für einzelne Aufgaben am Ende zu einer Note gelangen, ist Gegenstand der Interpretationsobjektivität. In diesem Zusammenhang spielen einerseits die gewählten Bezugsnormen eine große Rolle, aber möglicherweise auch die Entscheidungen von Fachkonferenzen: Vielleicht hat die Fachkonferenz zuvor festgelegt, dass ab 50 % erreichter Punktzahl eine 4,0 gegeben wird, ab 62,5 % eine 3,3, ab 75% eine 2,3 und ab 87,5 % eine 1,3. In diesem Falle ist zuvor festgelegt worden, wie wir von einer Punktzahl zur Note kommen.

In komplexeren diagnostischen Prozessen erkennt man den möglichen Einfluss der Interpretationsobjektivität auf die weitere schulische Laufbahn von Schülern und Schülerinnen. Nehmen wir an, dass das motorisch unruhige Verhalten eines Schülers von einer Lehrkraft als normale Begleiterscheinung von Wachstums- und Reifeprozessen interpretiert wird, von einer anderen Lehrkraft aber als fehlende Selbstregulation und Ausdruck des willentlichen Störens, so werden diese beiden Lehrkräfte sehr unterschiedlich auf das Verhalten des Schülers reagieren. Vielleicht liegen sogar beide Lehrkräfte falsch mit ihrer Interpretation des Verhaltens und nach einer tiefergehenden Diagnostik würde man zu dem Schluss kommen, dass ein Konzentrations- und Aufmerksamkeitsdefizit vorliegt. In diesem Fall würden beide zuvor genannten Interpretationen zu falschen Maßnahmen führen. Im ersten Fall würde vielleicht gar keine Maßnahme ergriffen werden, obwohl der Schüler Hilfe benötigt, im zweiten Fall würde der Schüler womöglich bestraft werden für ein Verhalten, welches er unter den jetzigen Bedingungen gar nicht steuern kann.

Beispiele zur Objektivität

Durchführungsobjektivität

Frau Klingel und Frau Süßholz diktieren in zwei Klassen derselben Jahrgangsstufe gleiche Diktat. Frau Klingel spricht langsam, lässt Pausen zwischen den Sätzen und wiederholt jeden Satzteil zweimal. Frau Süßholz spricht sehr leise, aber schnell und wiederholt jeden Satzteil nur einmal. In diesem Fall ist die Durchführungsobjektivität nicht gegeben.

Denn im ersten Fall haben die Schüler und Schülerinnen mehr Zeit zu schreiben und auch mehr Korrekturmöglichkeiten. Im zweiten Fall verstehen vor allem auch die hinten sitzenden Schüler und Schülerinnen teilweise Frau Süßholz falsch. Wir haben hier eine geringe Durchführungsobjektivität.

Auswertungsobjektivität

Die Lehrkräfte Frau Hinnerks und Herr Bauer entscheiden unabhängig voneinander wie viele Rechtschreibfehler in einem Diktat gemacht wurden und kommen zu derselben Anzahl von Fehlern. In diesem Fall ist die Auswertungsobjektivität hoch.

Würde aber Herr Bauer anders als Frau Hinnerks zu einer höheren Anzahl kommen, weil er beispielsweise sich über die Schrift von einzelnen Schülern ärgert und deshalb alles, was er als schlecht leserlich beurteilt zusätzlich als Rechtschreibfehler ankreidet, dann wäre die Auswertungsobjektivität eingeschränkt.

Interpretationsobjektivität

Katharina beteiligt sich kaum aktiv im Unterricht, stört diesen aber nicht. Frau Köhler und Frau Seibert, die als Klassenleitungsteam gemeinsam das Fach Deutsch unterrichten, machen beide unabhängig voneinander übereinstimmend diese Beobachtung. Frau Köhler interpretiert das Verhalten von Katharina als Desinteresse und beschließt ihr mündlich eine 3-4 zu geben. Frau Seibert interpretiert dagegen das Verhalten als introvertiert und aufmerksam und beurteilt dieses mit der Note 1-2. Die Interpretationsobjektivität ist deshalb gering, weil beide auf der Grundlage derselben Verhaltensbeobachtung zu einer unterschiedlichen Interpretation bezogen auf das zu erfassende Merkmal (mündliche Beteiligung) kommen.

Reliabilität

Die Zuverlässigkeit oder auch Reliabilität sagt uns, inwieweit ein Messinstrument das Merkmal, das es messen will, *genau* misst. Wenn Sie eine Küchenwaage haben und damit 500 g Mehl abwiegen, so erwarten sie, dass diese Messung unabhängig von der Tageszeit oder der Witterung ist. Sie nehmen an, dass die Waage immer genau das gleiche Gewicht anzeigt, ob sie morgens oder abends, bei gutem oder schlechtem Wetter ihre Portion Mehl abwiegen.

Die Höhe der Reliabilität wird mit einem Wert zwischen 0 und 1 angegeben. Je höher die Zuverlässigkeit, desto näher ist dieser Wert an 1. Wenn Sie eine Waage oder eine Uhr kaufen, gehen Sie davon aus, dass deren Zuverlässigkeit sehr nah an ‚1' liegt, sonst werden diese Geräte von Ihnen als untauglich eingeschätzt.

Um das Verhältnis von wahrem und verfälschendem Anteil zu schätzen, werden verschiedene Methoden verwendet. Bei der Testkonstruktion sind folgende Methoden gebräuchlich: Wiederholungs-, Halbierungs- und Paralleltestmethode. **Wichtig ist darauf hinzuweisen, dass diese Methoden ausschließlich dazu dienen, abzuschätzen, wie zuverlässig der Test ist, nicht dessen Zuverlässigkeit zu verbessern!**

Bleiben wir bei der Waage: Wenn Sie Ihre Packung Mehl wiegen und diesen Vorgang mehrfach hintereinander an verschiedenen Tagen und zu verschiedenen Tageszeiten wiederholen, dann sollte immer das gleiche Ergebnis herauskommen. Die Reliabilität wäre dann hoch. Und wir hätten die *Wiederholungsmethode* zur Abschätzung genutzt.

Bei Leistungsmessungen, Beobachtungen usw. ist das schwieriger. Wenn wir als Diagnostiker beispielsweise beobachten, dann sind wir selbst das Messinstrument und es stellt sich die Frage, ob wir in der Lage sind, tatsächlich bei einer wiederholten Beobachtung der gleichen Situation zur gleichen Einschätzung zu gelangen.

Nehmen wir dagegen ein objektives Testverfahren zur Erfassung von Depression, wo wir als Auswerter nicht mehr gefragt sind, weil die Antworten im Rahmen eines Multiple-Choice-Tests erhoben werden, dann müsste bei hoher Reliabilität das Ergebnis des ersten Durchlaufes mit dem Ergebnis der Wiederholung weitgehend übereinstimmen, da wir davon ausgehen, dass eine Depression über einen längeren Zeitraum *stabil* ist. Dennoch wird es Messfehler geben, da auch die Stimmung der Teilnehmer ihre Antworten teilweise beeinflussen wird.

Auch wenn wir also ein Testinstrument nehmen, das vielleicht sogar normiert wurde und dessen Zuverlässigkeit geprüft wurde, wie das beispielsweise bei Persönlichkeitstests und Intelligenztests der Fall ist, dann ist auch in diesem Fall davon auszugehen, dass das Instrument fehlerbehaftet ist, also keine hundertprozentige Zuverlässigkeit bietet. In diesem Fall wird ein sogenannter Standardmessfehler angegeben, der uns sagt *in welchem Bereich* der wahre Wert der Person, die den Test absolviert hat, liegen wird. Je höher die Reliabilität eines Tests ist, desto kleiner ist dieser Standardmessfehler. Nehmen wir an, in unserem IQ-Test läge dieser Standardmessfehler bei 5 Punkten und die Testperson hätte einen IQ-Wert von 128 Punkten erreicht. Der *wahre Wert* der Person läge dann zwischen 123 und 133. Damit könnten wir in diesem Falle also auch nicht ausschließen, dass diese Person vielleicht hochbegabt ist, da der obere Wert über der Grenze von 130 liegt, der Grenze für Hochbegabung. Die Person, die einen IQ-Wert oder einen anderen Testwert mitgeteilt bekommt, weiß also hinterher, dass ihr wahrer Wert in einem bestimmten Bereich um den erhobenen Wert liegen wird (vergl. Matheguru).

Auch bei den in der Schule eingesetzten Leistungstests wird es einen Messfehler geben, da davon auszugehen ist, dass die Zuverlässigkeit der Erhebung eingeschränkt ist. Unsere Leistungserfassung ist also grundsätzlich ebenfalls mit einem Messfehler behaftet.

Zusätzlich erschwerend kommt hinzu, dass anders als bei der Portion Mehl, die Stabilität der Merkmale (z.B. Argumentationsfähigkeit, Ausdrucksfähigkeit etc.), die wir messen wollen, gerade im pädagogischen Bereich nicht oder nur einschränkend gegeben ist.

Beispiel

Eine Lehrkraft erhebt mit einer Lernzielkontrolle, inwieweit die Schüler*innen das Prinzip der Schwerkraft verstanden haben. Dabei fällt ihr auf, dass bestimmte Fehler gehäuft auftreten. Sie vertieft daher nochmals mit den Schülern und Schülerinnen den Gegenstandsbereich. In der nachfolgenden Lernzielkontrolle schneiden die Schüler und Schülerinnen deutlich besser ab.

In diesem Beispiel sieht man, dass es gerade im Bereich des Lernens gar nicht das Ziel ist, überdauernde, stabile Fähigkeiten zu messen. Vielmehr sind wir an einer

fortlaufenden Entwicklung der Fähigkeiten und an einem Wissenszuwachs interessiert. Auch werden die Schüler, selbst wenn Sie keinen neuen Input mehr bekommen hätten, auch von der Lernzielkontrolle selbst gelernt und sich schon allein deshalb bei der zweiten Lernzielkontrolle verbessert haben.

Wo bleibt jetzt das Kriterium Reliabilität? Und was bedeutet dies dann noch im schulischen Kontext?

Bezogen auf das Beispiel können wir feststellen, dass es dennoch wichtig ist, möglichst zuverlässig den *momentanen* Leistungsstand zu erheben. Dennoch können wir zur Abschätzung der Zuverlässigkeit nicht die Wiederholungsmethode wählen wie wir gerade gesehen haben. Es gibt zwei weitere gängige Methoden, die sich besser eignen: Die Split-Half-reliabilität und die Paralleltestreliabilität.

Die *Paralleltestmethode* erzeugt zwei komplette Tests, die möglichst vergleichbar sind in Schwierigkeitsgrad, Länge und Inhalt. Auch wenn hier kein direkter Lernerfolg wie bei der Wiederholung derselben Arbeit gegeben ist, so muss dennoch davon ausgegangen werden, dass es im Falle der Bearbeitung der Tests zu zwei verschiedenen Zeitpunkten zu einem gewissen Lernzuwachs gekommen ist aufgrund der Erfahrung im ersten Test. Das ist aber wie schon gesagt, grundsätzlich im Rahmen schulischer Leistungserfassung auch erwünscht. Die Paralleltestmethode zeigt bei hoher Übereinstimmung der Ergebnisse (gemessen durch eine Korrelation der Testergebnisse), dass die Erfassung zuverlässig (reliabel) erfolgt ist.

Bei der *Split-Half-* oder auch *Halbierungsmethode* werden die Aufgaben eines Tests halbiert und getrennt ausgewertet, das Ergebnis beider Teile wird dann verglichen (miteinander korreliert). Auch hier gilt: Je höher die Übereinstimmung (oder auch je näher die Korrelation an 1), desto reliabler die Messung. Diese Methode eignet sich vor allem dann, wenn die Aufgaben ähnlich bzw. vergleichbar sind.

Im Rahmen eines Deutschaufsatzes wird es schwierig sein diese Methode anzuwenden, im Rahmen einer Mathematikarbeit zum Thema Bruchrechnen kann man sie eher einsetzen, indem man zum Beispiel die Ergebnisse der ungeraden Aufgaben (z.B. 1,3,5,7) mit den geraden Aufgaben (z.B. 2,4,6,8) vergleicht. Es ist darauf zu achten, dass möglichst nicht die erste Hälfte der Arbeit mit der zweiten Hälfte der Arbeit verglichen wird, da hier Zeit- und Ermüdungseffekte das Ergebnis verfälschen könnten: So steigt unter Umständen die Zahl der Fehler zum Ende der

Arbeit hin an und ebenso die Zahl der ungelösten Aufgaben, weil Schüler und Schülerinnen nicht mit der Arbeit fertig geworden sind.

Dennoch hat auch ein Vergleich der ersten und der zweiten Hälfte einer Arbeit einen Informationswert: Treten nämlich Ermüdungserscheinungen (Anstieg von Flüchtigkeitsfehlern) oder Zeitmangel auf, so muss sich die Lehrkraft fragen, ob sie mit dem Test tatsächlich noch gemessen hat, was sie zu messen beabsichtigte! Dazu mehr im nächsten Abschnitt zum Thema Validität.

Die Interne Konsistenz ist ein Maß, um abzuschätzen, inwiefern die gewählten Aufgaben oder Items in dieselbe Richtung zielen; also zu prüfen, ob die Items eines Tests beispielsweise tatsächlich alle das gleiche Merkmal messen. Dies ist insofern von Bedeutung als man nur unter dieser Voraussetzung die jeweiligen Werte, die eine Person auf einer Skala erreicht zusammenfassen und zum Beispiel aus diesen einen Skalenmittelwert berechnen darf. Das Verfahren ist ähnlich wie ein Paralleltest-Verfahren, nur dass hier jeweils die Korrelation *eines einzelnen Items* mit *allen anderen Items der Skala oder des Tests* betrachtet wird. Items, die nur gering mit der übrigen Skala korrelieren, sollten dann entfernt werden. Das Maß der Internen Konsistenz ist Cronbachs Alpha, je näher Cronbachs Alpha an „1“ ist, umso höher ist die interne Konsistenz.

Validität

Validität oder Gültigkeit beschreibt, ob mit dem gewählten Verfahren tatsächlich auch das gemessen wird, was zu messen beabsichtigt wurde. Es gibt unterschiedliche Arten der Validität, üblicherweise werden Inhaltsvalidität, Kriteriumsvalidität und Konstruktvalidität unterschieden (vergl. Arnold, 2001).

Inhaltsvalidität

Diese beschreibt inwieweit die Testinhalte repräsentativ sind und den gewählten (zu prüfenden) Inhaltsbereich tatsächlich abdecken. Spezialfälle der Inhaltsvalidität bilden die curriculare und die Unterrichtsvalidität. Bei der curricularen Validität geht es darum zu bestimmen, welche Aufgaben dazu geeignet sind, die wichtigsten Aspekte der laut Lehrplan zu lernenden Inhalte, abzudecken. Die curriculare Validität ist vor allem im Rahmen von großen Schulleistungstests relevant, da hier gemessen werden soll, inwieweit die Anforderungen des Curriculums auf schulischer Ebene erfüllt worden sind. Welche

Aufgaben letztlich dazu geeignet sind, die wichtigsten Aspekte zu erfassen, ist in der Regel strittig und wird von den Experten kontrovers diskutiert (vergl. Rindermann, 2008).

Dabei ist zu beachten, dass es natürlich einen Unterschied zwischen den Anforderungen, die im Curriculum aufgeführt sind und der spezifischen Ausgestaltung des Curriculums auf Schul- und Klassenebene gibt: Welche Bücher werden von der Lehrkraft benutzt? Wie gestaltet sie den Lernprozess? Welche Schwerpunkte setzt sie? Diese Entscheidungen beeinflussen maßgeblich das, was tatsächlich gelernt wird. Man unterscheidet daher zwischen intendiertem Curriculum, das im Rahmen von bildungspolitischen Entscheidungen festgelegt und in vergleichenden Studien geprüft werden kann und dem realisierten Curriculum, der tatsächlichen Umsetzung des Curriculums im Unterricht. Entsprechend stellt sich für die Lehrkraft bei der Gestaltung der Leistungsüberprüfung weniger die Frage nach der allgemeinen curricularen Gültigkeit als vielmehr nach der *Unterrichtsvalidität*: Bildet der Test tatsächlich das ab, was auch Gegenstand des Unterrichts war? Oder geht er darüber hinaus? Dies ist insofern bedeutsam als man argumentieren kann, dass der Unterricht jedem gleichermaßen die Gelegenheit zum Lernen geben muss und dass auch nur das, was tatsächlich Lerngegenstand war, geprüft werden sollte. Ansonsten liefe man Gefahr, dass die Einflüsse anderer Faktoren für das Abschneiden der Schüler und Schülerinnen bestimmend werden: z.B. deren Bildungshintergrund, das Engagement und die Unterstützung der Eltern (oder der Nachhilfe) beim Lernen, das eigene Interesse der Schüler und Schülerinnen für bestimmte Lerngegenstände.

Kriteriumsvalidität

Die Kriteriumsvalidität sagt uns, inwieweit unser Testergebnis mit Testverfahren, die das Gleiche zu messen beanspruchen, übereinstimmt (Übereinstimmungsvalidität oder auch konvergente Validität) oder sich von anderen unterscheidet (diskriminante Validität). Das ist insofern bedeutsam, als dass ein Testverfahren einerseits natürlich mit Verfahren, die Gleiches messen, eine große Übereinstimmung aufweisen sollte, es sich hingegen von anderen Phänomenen oder Konstrukten abgrenzen muss.

Beispiel

Nehmen wir an, Sie befragen mit Hilfe eines Fragebogens verschiedene Lehrkräfte inwieweit Schüler und Schülerinnen einer Klasse als hyperaktiv wahrgenommen werden. Sie stellen fest, dass die Lehrkräfte eine hohe Übereinstimmung in ihrem Urteil aufweisen und alle zu einem vergleichbaren Urteil gekommen sind.

Jetzt engagieren Sie unabhängige Beobachter, die anhand eines diagnostischen Beobachtungsbogens, der ebenfalls den Anspruch hat Hyperaktivität zu erfassen, systematisch über viele Unterrichtsstunden die Klasse beobachten und die Schüler und Schülerinnen einschätzen.

Am Ende vergleichen Sie beide Urteile miteinander – die aus den Beobachtungen und die aus der Fragebogenerhebung. Sie stellen eine hohe Übereinstimmung fest.

Unter der Annahme, dass die vorherigen Gütekriterien der Reliabilität und Objektivität in unserem Beispiel in hohem Maße erfüllt worden sind (das ist wichtig!!!), sollten die Ergebnisse der Beobachtung mit den Ergebnissen unseres Fragebogens eine hohe Übereinstimmung aufweisen, weil sie ja beide das Gleiche messen.

Ist das der Fall, dann ist die Übereinstimmungsvalidität hoch! Andererseits sollte unser Fragebogen aber auch eine niedrige Übereinstimmung haben mit zum Beispiel einem Fragebogen, der erhebt inwieweit die Schüler den Unterricht dadurch stören, dass sie sich mit anderen Dingen oder Personen beschäftigen. Ist das wiederum der Fall, dann wäre die *diskriminante Validität* hoch, oder anders ausgedrückt: Störverhalten könnte unabhängig davon erfasst werden, ob die Person hyperaktiv ist oder nicht. Es gibt sicherlich Fälle, in denen sowohl Hyperaktivität als auch Störverhalten auftreten. Wenn wir jedoch eine große Gruppe von Schülern und Schülerinnen beobachtet und eingeschätzt (Fragebogen) haben, dann sollten sich beide Merkmale unterschiedlich verteilen: D.h. es gäbe keinen systematischen Zusammenhang zwischen beiden Merkmalen. Oder anders ausgedrückt für unser Beispiel: Es gibt Schüler und Schülerinnen, die

stören ohne irgendwelche Anzeichen von Hyperaktivität aufzuweisen, genauso wie es störende und nicht störende Schüler und Schülerinnen gibt, die Hyperaktivität zeigen.

Auch die Vorhersagevalidität oder auch prognostische Validität gehört zur Kriteriumsvalidität. Im Unterschied zur eben beschriebenen Übereinstimmungsvalidität liegt hier das Vergleichskriterium in der Zukunft. Wenn ich also beispielsweise einen Test vorliegen habe, der vorgibt, Studienerfolg vorhersagen zu können, dann sollte dieser mit dem tatsächlich in späteren Jahren feststellbarem Studienerfolg hoch korrelieren oder anders gesagt in hohem Maße übereinstimmen.

Konstruktvalidität

Im Rahmen der Konstruktvalidität geht es darum zu prüfen, inwieweit die Ergebnisse eines Verfahrens oder Tests das Konstrukt tatsächlich erfassen. Die Konstruktvalidierung befasst sich also mit der theoretischen Klärung der Frage, was der Test misst (vergl. Lienert & Raatz, 1998). Ein Konstrukt kann nicht direkt beobachtet werden. Es wird abgeleitet aus dem, was wir beobachten oder erfassen (z.B. durch Antworten in einem Fragebogen oder durch die Leistung in einem Test). Die Prüfung der Konstruktvalidität ist in der Regel sehr aufwendig. Man nutzt beispielsweise Faktorenanalysen, um herauszufinden welche Teile des Tests miteinander korrelieren, also eine hohe Übereinstimmung aufweisen. Dabei wird immer der Bezug zu einem theoretischen Modell hergestellt.

Nehmen wir an, Sie möchten das Konstrukt „Prüfungsangst“ erfassen. Sie nutzen dazu einen Fragebogen, der die Schüler und Schülerinnen auffordert, ihr Empfinden in spezifischen Prüfungssituationen einzuschätzen. Der Fragebogen sollte dann zum Beispiel unterschiedliche Dimensionen von Prüfungsangst abbilden, wenn er sich an einem theoretischen Modell von Prüfungsangst orientiert: so wäre es unsinnig nur nach körperlichen Empfindungen wie Aufregung, erhöhter Herzschlag, Schweißausbrüchen etc. zu fragen, gleichzeitig aber z.B. kognitive Faktoren wie begleitende Gedanken und Einstellung zur Prüfung (z.B. „Ich mache mir vor der Prüfung große Sorgen, ob ich die Prüfung schaffen kann.“) auszublenden. Sie müssten also auf der Grundlage des theoretischen Konstrukts „Prüfungsangst“ die unterschiedlichen Dimensionen der Prüfungsangst sinnvoll operationalisieren und somit messbar machen. Das bedeutet, Sie müssten sich überlegen wie diese Dimensionen erfasst werden können.

Beispiel

Sie wollen das Konstrukt „Kommunikationskompetenz“ erfassen und als eine Dimension dieser Kompetenz, die Fähigkeit auf den Gesprächspartner und dessen Beiträge einzugehen, erheben. Wie können Sie diesen Teil von Kommunikationskompetenz operationalisieren?

Zunächst wäre die Frage, ob eine einfache Befragung im Sinne einer Selbsteinschätzung „Ich gehe in Diskussionen auf die Beiträge meines Diskussionspartners ein.“ sinnvoll ist.

Dies ist womöglich schwierig, weil erstens diese Aussage eine hohe Erwünschtheit hat, d.h. die meisten von uns würden dem zustimmen wollen, zweitens sich die Frage stellt, was der oder die Einzelne denn unter „auf Beiträge eingehen“ überhaupt versteht.

Wir müssen uns also zunächst fragen, was „eingehen auf Beiträge anderer?“ bedeutet:

Zunächst könnte eine Person sich in der Antwort auf den Vorredner beziehen. Das könnte aber sehr unterschiedlich aussehen: Vielleicht nutzt der Redner die Aussage des Vorredners nur im Sinne eines Sprungbretts, um eigene Gedanken auszuführen, vernachlässigt aber das vom Vorredner angeführte Argument. Oder der Redner greift das Argument auf, um es zu diskreditieren. Hier hätten wir uns zu fragen, ob wir das noch als „eingehen auf den Redebeitrag des Diskussionspartners“ werten wollen, oder ob es hier schon um eine andere Facette geht, nämlich „sachlicher Umgang mit Gegenargumenten“. Hätten wir diese Fragen geklärt, müssten wir uns schließlich überlegen, ob wir insgesamt auch alle Facetten dieser Dimension hinreichend erfasst hätten und auch, wie wir diese Facetten jeweils gewichten wollen. Das ist vor allem dann entscheidend, wenn wir im Abschluss zu einem Urteil über das Erreichen dieser Kompetenz gelangen wollen!

Abschließend kann man festhalten, dass im schulischen Alltag vor allem Inhaltsvalidität und Übereinstimmungs- und Vorhersagevalidität eine große Bedeutung haben. Die Konstruktvalidität findet vor allem in Schulleistungstests Beachtung, die ja zum Teil den Anspruch erheben, spezifische Kompetenzen zu

erheben. Das bedeutet nicht, dass Konstruktvalidität für die Lehrkraft irrelevant ist, aber im schulischen Alltag werden häufig ausgewählte Dimensionen von Kompetenzen anhand von spezifischen Aufgabenformaten und Unterrichtsmaterialien gelehrt und entsprechend überprüft.

EINHALTUNG UND VERBESSERUNG VON GÜTEKRITERIEN

Im folgenden Kapitel geht es darum zu prüfen, welche Möglichkeiten es gibt die Gütekriterien zu verbessern. Dabei ist es wichtig die verschiedenen Beurteilungsfehler kennenzulernen und zu diskutieren wie diese vermindert werden können. Wir beginnen mit der Durchführungsobjektivität. Denn zu Beginn einer Diagnostik steht die Absicht, etwas zu betrachten oder zu erheben und diese Erhebung durchzuführen. Daher lautet die erste Frage:

Welche Möglichkeiten gibt es, die Durchführungsobjektivität zu erhöhen? Zunächst kann man versuchen, möglichst faire Testbedingungen für alle herzustellen. Das heißt nicht, dass diese für alle Personen gleich sein müssen.

So mag zum Beispiel eine Schülerin, die sich schwer konzentrieren kann, von einem Einzelplatz profitieren, während ein ängstlicher Schüler dadurch stark verunsichert wird. Es gilt also für die Durchführung des Verfahrens sicherzustellen, dass alle Personen möglichst in der Lage sind, zu zeigen, was sie können bzw. zu zeigen wie sie sich in der Regel in bestimmten Situationen verhalten (z.B. im Rahmen einer Verhaltensbeobachtung). Als Prüfer*in oder Diagnostiker*in sollte man motivierend und positiv wirken ohne manipulativ in den Prozess einzugreifen.

Man kann in Prüfungen Anweisungen standardisieren, sich vorher genau überlegen wie man in die Testsituation einführen möchte. Es ist auch möglich, Rückfragen öffentlich zuzulassen, sodass man versucht sicherzustellen, dass alle verstanden haben, was sie zu tun haben. Weiter sollte das eigene Verhalten neutral bzw. positiv gestimmt sein. Es ist ungünstig herumzugehen, Schülern und Schülerinnen über die Schulter zu blicken und dazu noch Kommentare zu machen. Auch störende Beschäftigungen (Spiele auf Tablets, Klackern der Tastatur, lautes hin- und herlaufen) sollten vermieden werden.

Letztlich geht es bei der Durchführungsobjektivität darum, negative Einflüsse zu minimieren: Seien es Störungen von außen, Missverständnisse, Prüfungsangst oder Ablenkungen. Dem zugrunde liegt die Annahme, dass negative Einflüsse die Wahrscheinlichkeit erhöhen, dass wir etwas Anderes messen als das, was wir messen wollen. Für die Erfassung von Leistungen z.B. wäre eine große Angst vor Prüfer*innen damit verbunden, dass die Person ihre Leistungsfähigkeit gar nicht zeigen *kann*, da diese von der Angst überlagert wird. Oder wären die Teilnehmer*innen einer PISA-Erhebung zum Beispiel nicht motiviert an der

Untersuchung teilzunehmen, dann müssen wir fürchten, dass sie ihr Potential möglicherweise nicht zeigen, weil sie mit unterdurchschnittlichem Engagement die Aufgaben bearbeiten oder schwierige Aufgaben gleich weglassen, da sie zu anstrengend sind. In der Folge hätten wir dann aber eher die Anstrengungsbereitschaft der Schüler und Schülerinnen gemessen und nicht ihre Kompetenz. Hier würde es also darum gehen, zunächst die Schüler und Schülerinnen dazu zu motivieren, wirklich ihr Bestes zu geben.

Bei der Auswertungsobjektivität liegt die Herausforderung darin, die möglichen **Beurteilungsfehler** zu reflektieren und diese, wenn es geht, zu vermeiden. Es gibt eine Vielzahl von Beurteilungsfehlern. Einige entstehen aufgrund der Mehrdeutigkeit von Situationen oder Begriffen, andere durch mangelnde Sorgfalt, wieder andere dadurch, dass unser Gedächtnis uns einen Streich spielt.

Mehrdeutigkeit

Wenn wir zum Beispiel die erstellte Mindmap eines Schülers beurteilen wollen, dann werden wir in Abhängigkeit davon, wie sehr sie mit unserer eigenen Vorstellung vom Gegenstand übereinstimmt, mehr oder weniger geneigt sein, diese als „korrekt“ oder „erschöpfend“ wahrzunehmen. Damit ist gemeint, dass verschiedene Menschen zu einem Begriff Unterschiedliches assoziieren werden. Was sie assoziieren hängt unter anderem von ihren Vorerfahrungen, ihrem Vorwissen und ihrer Sozialisation ab. Zudem belegen wir bestimmte Begriffe auch mit Gefühlen der Zustimmung oder Ablehnung oder mit einer unterschiedlichen Intensität. So mag eine Person eine mittlere Ablehnung ausdrücken, wenn sie sagt: „Ich hasse Eiscreme.“ Während für eine andere Person diese Aussage bedeuten würde, dass sie eine derart starke Abneigung verspürt, dass ihr schon beim Anblick von Eiscreme, schlecht wird. Das Wort „hassen“ wird in beiden Fällen anderes verwendet und interpretiert.

Auch fachliche Begriffe sind häufig unscharf. Selbst wenn in dem Deutschbuch steht, was verlangt wird, wenn in der Aufgabe „erörtere…“ steht, eröffnet die Anweisung immer noch einen gewissen Interpretationsspielraum. Diejenigen, die die Aufgabe bearbeiten, interpretieren die Anweisung unterschiedlich genauso wie diejenigen, die die Bearbeitung nachfolgend auswerten.

Kann man aus diesem Dilemma herauskommen, wenn man etwas beobachten oder etwas Geschriebenes interpretieren möchte?

Die Antwort ist, dass man versuchen kann die Mehrdeutigkeit zu reduzieren, indem man beispielsweise nachfragt, was denn genau gemeint war. Aber auch, indem man genauer beschreibt, was man wirklich erwartet und durch Rückfragen sich versichert, ob dies auch alle so verstanden haben.

In der Leistungsüberprüfung spielt die Erstellung eines Erwartungshorizonts eine besondere Rolle: Der Erwartungshorizont bietet sowohl den Prüflingen als auch den Prüfer*innen einen Orientierungsrahmen und macht, wenn er sinnvoll differenziert und klar gestaltet ist, transparent, was gefordert ist und was nicht. Die Qualität des Erwartungshorizonts kann auch die Gütekriterien der Reliabilität und der Validität beeinflussen: Werden in den Erwartungen auch Fähigkeiten oder Dinge erfasst, die gar nicht Gegenstand der Prüfung sein sollten? Und bezogen auf die Reliabilität nehmen wir an, dass die Beurteilung anhand eines Erwartungshorizonts die Wahrscheinlichkeit erhöht, dass dieselben Prüfer*innen die Arbeit auch bei der zweiten Korrektur gleich beurteilen.

Im Rahmen von Beobachtungen kann man sogenannte Beobachterschulungen vornehmen. Vorab wird überlegt, welche Verhaltensweisen überhaupt relevant sind für das, was man beobachten möchte. Man operationalisiert also das, was man prüfen will, d.h. man versucht möglichst viele der möglichen Verhaltensweisen zu erfassen, die auf das zu prüfende Konstrukt Rückschlüsse erlauben könnten. Mithilfe von Beobachtungsrastern übt man das Beobachten und lernt über die Differenzen (mehrere Beobachter*innen haben unterschiedliches beobachtet) wie groß der Interpretationsspielraum tatsächlich ist. Auch hier hat der Einsatz von Beobachtungsrastern das Ziel, die Reliabilität zu erhöhen, damit wir auch bei einer erneuten Beobachtung zu gleichen Einschätzungen kommen.

Beispiel

Sie wollen im Rahmen Ihres Praktikums beobachten, inwieweit Schüler und Schülerinnen aufmerksam das Unterrichtsgeschehen verfolgen.

Dazu entwickeln Sie ein Beobachtungsraster, das unterschiedliche Aspekte der Aufmerksamkeit operationalisiert. So gibt es z.B. den Bereich Blickrichtung, in dem Sie eintragen können, ob Schüler X den Blick auf das Zentrum des Unterrichtsgeschehens (Lehrkraft, Tafel, Mitschüler, die Beiträge leisten etc.) richtet oder nicht. Jetzt müssten Sie aber noch überlegen, ob tatsächlich alle anderen Blickrichtungen fehlende Aufmerksamkeit bedeuten, oder ob Sie vielleicht differenzieren wollen zwischen dem hilfesuchenden Blick im Buch oder zum Nachbarn. Was uns wiederum zur nächsten Frage führen wird: Wie unterschiedlich kann ein Blick sein: Fragend, abwesend, gelangweilt, interessiert - und woran entscheiden wir, wie wir diesen Blick zuordnen?

Nehmen wir an, Sie würden all diese Fragen für sich klären können, dann wäre doch abschließend zu überlegen, ob mit Hilfe Ihres Beobachtungsrasters verschiedene Beobachter*innen zu vergleichbaren Ergebnissen kommen.

Ist dies nämlich nicht der Fall, könnten wir anhand der Unterschiede und Gespräche herausarbeiten, wo die Kategorien uneindeutig oder unscharf sind. Daraufhin könnten Sie das Beobachtungsraster verbessern, differenzieren etc. und einen weiteren Versuch durchführen, um zu testen, ob sich das Instrument nun bewährt.

Gedächtnisfehler

Ein typischer Gedächtnisfehler (Jäger, 2007) ist das Fehlen von Erinnerung an bestimmte Sachverhalte, weil zum Beispiel diese von anderen aufregenderen Ereignissen überlagert wurden, weil starke Empfindungen wie große Angst oder Stress dazu führen, dass wir uns nicht mehr erinnern bzw. nur an bestimmte Dinge erinnern. So kann Stress auch dazu führen, dass wir uns auf einen bestimmten Gegenstand konzentrieren und andere Dinge, die gleichzeitig passieren, ausblenden und folglich auch nicht mehr erinnern. Andere Gedächtnisfehler sind das Durcheinanderbringen von zeitlichen Abfolgen oder die Tendenz, den

Ereignissen, die für uns persönlich bedeutsamer waren, größeres Gewicht beizumessen. So wird eine Schülerin, die in einem Streit massiv persönlich beleidigt wird, dies in der Intensität anders wahrnehmen als eine dabeistehende Lehrkraft, die vielleicht die Überzeugung hat, dass das doch nur etwas derbe Ausrutscher von Teenagern waren, die das gar nicht so gemeint haben.

Es ist schwierig hier von Wahrheit oder Unwahrheit zu sprechen, da wir normalerweise keine Aufzeichnungen haben, anhand derer wir gemeinsam das Geschehen rekonstruieren und besprechen könnten. Ein Weg wäre hier zunächst zu akzeptieren, dass es unterschiedliche Wahrnehmungen und Erinnerungen gibt, ohne diese abzuwerten.

Erschwert werden kann die Beurteilung noch dadurch, dass Beobachter*innen oder Beurteiler*innen nicht sorgfältig vorgeht oder unaufmerksam ist. Daraus könnten Fehler entstehen, die unabsichtlich sind, aber dennoch die spätere Beurteilung deutlich beeinflussen.

Bezogen auf die genannten Probleme hilft sicherlich auch eine kritische Reflexion des eigenen Vorgehens: Habe ich mich wirklich bemüht, alle relevanten Aspekte wahrzunehmen und zu würdigen? Wo müsste ich vielleicht noch Informationen sammeln, um zu einem korrekten Urteil zu gelangen?

Generalisierungsfehler

Eine weitere Art von Fehlern ist die, von bestimmten Ereignissen oder Beobachtungen auf das folgende Verhalten oder auf andere nicht beobachtbare Eigenschaften zu schließen.

Der *Haloeffekt* (Jäger, 2007) zum Beispiel beschreibt, dass die Wahrnehmung einer bestimmten Eigenschaft einer Person alle anderen Eigenschaften ‚überstrahlt'. So könnten wir eine Person, die wir als menschlich sehr engagiert und selbstlos erleben, primär mit dieser ‚Brille' betrachten, und alle ihre Handlungen unter eben dieser Prämisse ‚sie will etwas Gutes tun' wahrnehmen und dabei andere Motive oder Eigenschaften übersehen.

Auch Persönlichkeitstheorien, die wir haben, können diese Effekte erzeugen. Wir denken vielleicht, dass jemand, der hochbegabt ist, sozial inkompetent sein wird. Wenn wir implizit diese Theorie verinnerlicht haben, wird dies unsere Wahrnehmung entsprechend einfärben.

Der *Ankereffekt* besagt, dass je nachdem welche Grundannahme wir über etwas haben, wir Dinge unterschiedlich betrachten: So werden wir unter der Annahme, „Schüler und Schülerinnen werden bei einer Klassenarbeit immer versuchen zu schummeln“, den Blick eines Schülers zu seiner Nachbarin anders interpretieren als wenn wir diese Annahme nicht teilen.

Die *Tendenz eine konsistente Darstellung* abzugeben entspricht einem weiteren Fehler: Hier versuchen wir, zu einem runden Gesamturteil zu kommen. Wir neigen dazu, die Informationen, die eigentlich nicht ins Bild passen zu vernachlässigen, damit das Bild in sich stimmig bleibt.

Letztlich gilt bei diesen Fehlern ähnlich wie bei den vorigen, dass das Wissen um diese Probleme und die Bereitschaft die eigenen Urteile und Wahrnehmungen kritisch zu hinterfragen, die beste Möglichkeit darstellen, diesen Fehlern vorzubeugen (Jäger, 2007; Stemmler & Margraf-Stiksrud, 2015).

Wahrnehmungsfehler, die durch das Verhalten der Beobachteten oder Beurteilten entstehen

Auch das Wissen, Gegenstand einer Beurteilung oder Beobachtung zu werden, kann das gezeigte Verhalten derart verändern, dass die Beobachtung verfälscht wird.

So lernen auch Schüler und Schülerinnen, welche Antworten von ihnen erwartet werden, also *sozial erwünscht* sind. Das bedeutet auch, dass sie unter Umständen ihr „wahres“ Verhalten zu bestimmten Zeitpunkten verschleiern können. Auch in Bezug auf vergangene Sachverhalte besteht die Möglichkeit, zu *lügen* oder die Situation zu *bagatellisieren*. Eine weitere Verfälschungsmöglichkeit ist das *Simulieren* als Vortäuschen von etwas, das gar nicht so stattgefunden hat.

Beispiel

Nehmen wir an, die Schülerin Melanie kommt zu ihnen und beschwert sich, weil sie von ihrer Mitschülerin Julia massiv bedroht worden ist. Sie kennen Julia als leistungsstark und erleben sie im Umgang mit Lehrkräften und im Unterricht als höflich und angepasst.

Sie bitten Julia dazu und fragen sie, was vorgefallen ist. Julia weiß, was von ihr erwartet wird und verschleiert absichtlich und durchaus wortgewandt ihr Verhalten und ihre Motive. Melanie dagegen, die sprachlich nicht so gewandt ist, ist nicht in der Lage dem etwas entgegenzusetzen.

In der oben geschilderten Situation können sich zwei Fehler gegenseitig verstärken: Einerseits die Tendenz zur unkritischen Zustimmung der Lehrkraft zu Julias Verhalten, weil sie Julia grundsätzlich als *sozial kompetent* einschätzt und andererseits Julias absichtliche Verfälschung des Geschehenen. Die Wahrnehmung von Julia als ‚sozial kompetent' könnte zu einer *selektiven Wahrnehmung* der Situation führen: Sie würden dann primär Anzeichen, die *für* die soziale Kompetenz von Julia sprechen registrieren, widersprüchliche Anzeichen *dagegen* ignorieren.

Für Lehrkräfte ist vor allem die Befassung mit Konflikten, denen sie selbst nicht beigewohnt hat, schwierig. Einerseits wollen sie den Konflikt fair klären und natürlich erstmal davon ausgehen können, dass sie ‚die Wahrheit' erzählt bekommen, andererseits gibt es natürlich auch unabsichtliche Verzerrungstendenzen. Auch Schüler und Schülerinnen, die Zeugen eines Konflikts waren, unterliegen bei der Beobachtung den gleichen Beobachtungsfehlern wie die Lehrkraft selbst.

Wir haben hier also ein kaum aufzulösendes Dilemma. Letztlich besteht nur die Möglichkeit, zunächst *alle* Sichtweisen möglichst unvoreingenommen zur Kenntnis zu nehmen und diese anzuerkennen. Im zweiten Schritt sollte dann durch gezieltes Nachfragen und Vermittlung zwischen den Positionen eine Konfliktlösung erzielt werden.

Dabei ist darauf zu achten, die eigenen Vorannahmen kritisch zu reflektieren. Darüber hinaus sollte darauf hingearbeitet werden, dass in der Klasse ein

Klassenklima entsteht, das von Ehrlichkeit, Fehlertoleranz und Offenheit geprägt ist. Werden Fehler massiv bestraft, wird das eher dazu führen, dass Fehlverhalten verdeckt wird, anstatt dass es sinnvoll bearbeitet wird. Zeigt die Lehrkraft dagegen durch ihr eigenes Verhalten, dass sie Fehler eingesteht und bereit ist, Kritik sachlich anzunehmen und ihr Verhalten zu verändern, dann steigt die Wahrscheinlichkeit, dass die Schüler und Schülerinnen sich ebenso versuchen so zu verhalten.

Fehler, die durch Reihenfolgeeffekte oder Urteilstendenzen entstehen

Eine Fehlerkategorie sind Fehler, die entweder dadurch in welcher Reihenfolge eine Beurteilung erfolgt oder aufgrund von Urteilstendenzen entstehen (Jäger, 2007; Stemmler & Margraf-Stiksrud, 2015). So könnte die Wahrnehmung eines Aufsatzes davon beeinflusst werden, wie der *zuvor* gelesene Aufsatz eingeschätzt wurde (*Positions- und serialer Effekt*). War dieser besonders gut, dann besteht die Gefahr, dass wir einen nachfolgenden Aufsatz, der objektiv etwas schlechter ist, strenger beurteilen als wenn von uns zuvor ein Aufsatz von vergleichbarer Qualität gelesen wurde.

Hier kann es hilfreich sein, Arbeiten zweimal in unterschiedlicher Reihenfolge zu korrigieren, dies trifft vor allem auf Arbeiten zu, deren Antworten sich nicht einfach als richtig oder falsch einordnen lassen.

Darüber hinaus unterscheiden wir uns auch alle darin, wie wir generell geneigt sind Urteile vorzunehmen: So könnte eine Person eher zu *extremen Urteilen* neigen (also etwas als z.B. besonders gut oder besonders schlecht einschätzen), während eine andere Person eine *Tendenz zur Mitte* hat. Letztere nimmt eher auch gegensätzliche Merkmale wahr und wird dann in einer abschließenden Beurteilung zu einem ‚weder noch' tendieren, d.h. sie wird eher eine mittlere Einschätzung vornehmen, weil sie sich nicht für das eine oder andere entscheiden kann.

Beispiel

Frau Konhorst schätzt die mündliche Beteiligung ihrer Englischklasse ein und neigt zu extremen Beurteilungen:

Sie gibt den stillen Schülern und Schülerinnen eine deutlich schlechtere Beurteilung als den Schülern und Schülerinnen, die sich stark beteiligt haben. Diese schätzt sie durchgängig als ‚sehr gut' ein, unabhängig von der Qualität ihrer Beiträge. Die stillen Schüler und Schülerinnen bekommen dagegen Noten im Bereich ausreichend bis mangelhaft. Die Notenverteilung zeigt, dass viele Schüler Noten in den Extrembereichen haben.

Herr Dierks, der die Klasse zuvor unterrichtete, kam aufgrund seiner Tendenz zur Mitte zu einer anderen Einschätzung. Er vermutete aufgrund anderer Ereignisse (Tests, Stillarbeit), dass die stilleren Schüler und Schülerinnen sich auf andere Art beteiligten und sieht ihre mündliche Mitarbeit ambivalent, ebenso schätzt er die starke mündliche Beteiligung nicht nur als positiv ein, weil einige der Schüler und Schülerinnen auch häufiger fehlerhafte Antworten gegeben haben. Das führt dazu, dass in seiner Notenverteilung die meisten Schüler und Schülerinnen zwischen gut und befriedigend stehen und es nur wenige Noten in den Randbereichen gibt.

Wir sehen an diesem Beispiel nicht nur die unterschiedlichen Beurteilungstendenzen, sondern damit einhergehend auch eine unterschiedliche Operationalisierung der Frage, was eine „gute mündliche Beteiligung" ausmacht. Während Frau Konhorst sich primär an der tatsächlichen Unterrichtsbeteiligung orientiert, fließen in die Bewertung von Herrn Dierks noch ganz andere Faktoren ein, die streng genommen auch anderes messen: Lernverhalten (Vokabeltests) und Qualität der Stillarbeit.

Weitere Beurteilungstendenzen, die sich negativ auf die objektive Einschätzung auswirken können, sind die Tendenz zur unkritischen Zustimmung oder unkritischen Ablehnung, die Tendenz zu undifferenzierten Beobachtungen und Beurteilungen, sowie Ähnlichkeits- und Kontrasteffekte und der Pygmalioneffekt (vergl. Jäger, 2007).

Bei der unkritischen Zustimmung, hinterfragen wir nicht das Verhalten, das wir beobachten (z.B. eine geschönte Selbstdarstellung wie im Beispiel weiter oben, wir akzeptieren dies als „wahr"), bei der Tendenz zur unkritischen Ablehnung,

lehnen wir ab, was wir wahrnehmen, weil es z.B. möglicherweise nicht in das Bild passt, was wir sonst von der Person haben. Wir kommen vielleicht zu dem Urteil „Das kann ja gar nicht sein.“, dadurch schätzen wir die Situation falsch ein.

Ein weiterer Fehler kann darin liegen, dass wir das, was wir beobachten wollen, nicht genügend differenzieren. Die beobachteten Merkmale verschwimmen zu einem Gesamteindruck.

Beispiel

Herr Lehrters beobachtet den vierzehnjährigen Leon im Unterricht, der manchmal mit seinem Nachbarn quatscht, sich dann aber wieder aktiv am Unterricht beteiligt. Auch fällt ihm auf, dass die Mappen von Leon an den Ecken geknickt sind und einzelne Blätter Tintenflecken aufweisen. Da Leon in den schriftlichen Leistungsprüfungen meist sehr gut abschneidet, auch wenn Herr Lehrters vom häufigen Einsatz des Tintenkillers genervt ist, beauftragt Herr Lehrters Leon häufig damit, seine Sitznachbarn in der Stillarbeit zu unterstützen. All diese Eindrücke verschwimmen am Ende des Jahres zu dem Urteil, dass das Arbeitsverhalten von Leon deutlich verbesserungswürdig sei.

Am oben genannten Beispiel kann man sehen, dass Herr Lehrters einzelne Merkmale relativ unreflektiert zu einem Gesamturteil verschmelzen lässt, ohne genauer zu differenzieren, was er genau beobachtet hat. So könnte der Hinweis darauf, dass er sich eine bessere Mappenführung oder saubereres Arbeiten wünscht, auch dahingehend hinterfragt werden, ob der häufige Einsatz des Tintenkillers in Arbeiten ein Hinweis darauf ist, dass Leon seine Antworten in den Tests überdenkt und nochmals korrigiert und genau deshalb dort gut abschneidet. Weiter wäre zu fragen, worüber Leon überhaupt mit seinem Sitznachbarn redet, ist es immer „stören“ oder nimmt er teilweise den Auftrag, sich um seine Sitznachbarn zu „kümmern“ wahr und klärt im Gespräch deren Fragen?

Welche Merkmale gehören überhaupt zu dem Konstrukt „Arbeitsverhalten“ und wie genau gehen diese in eine abschließende Beurteilung ein?

Man könnte hier noch weiter ausführen, dass es sicherlich interessant sein könnte mit Leon zu sprechen und zu fragen wie er die ihm zugewiesene Rolle (andere zu unterstützen) wahrnimmt und wie er sich selbst sieht.

Ein weiterer Beurteilungseffekt ist der *Ähnlichkeits- oder Kontrasteffekt*. Das bedeutet, dass wir diesen Fehler dann begehen, wenn wir die zu beurteilende Person anders beurteilen, weil sie uns *ähnlich* ist, oder weil wir ihr Verhalten *im Kontrast* zu unseren eigenen Persönlichkeitsmerkmalen sehen.

Der *Pygmalioneffekt* schließlich ähnelt dem Effekt der sozialen Erwünschtheit. Hier senden der/die Untersucher/Beobachter*in oder Prüfer*in subtile Signale aus, welche Ergebnisse er oder sie für wünschenswert hält und die Prüflinge versuchen sich entsprechend zu verhalten.

Abschließend kann man festhalten, dass unsere Einschätzungen alle einer Reihe von Beurteilungsfehlern und -tendenzen unterliegen. Diese lassen sich wahrscheinlich niemals alle kontrollieren, aber alle Fehler lassen sich zumindest durch Reflexion des eigenen Verhaltens und der eigenen Beurteilungen reduzieren und unter Umständen auch revidieren. Die Aufgabe, das eigene Verhalten immer wieder zu reflektieren und zu hinterfragen begleitet uns als Lehrkraft durchgängig über unser ganzes Berufsleben und kann niemals abschließend bearbeitet werden. Das liegt auch daran, dass sich Schüler und Schülerinnen im Laufe der Zeit verändern. Auch wenn uns Routinen und Erfahrungen dabei helfen, Probleme zu erkennen und Situationen einzuschätzen, so bergen sie immer auch die Gefahr, dass wir uns auf alte Erfahrungen verlassen und uns alternativen Erklärungsansätzen gegenüber verschließen. Dieser Gefahr können wir nur dadurch begegnen, dass wir unser Vorgehen immer wieder aufs Neue überprüfen und hinterfragen.

SCHULLEISTUNGEN DIAGNOSTIZIEREN

Im nun folgenden Kapitel beschäftigen wir uns mit der Erfassung von Schulleistungen. In diesem Zusammenhang muss zunächst geklärt werden, was unter Schulleistung zu verstehen ist. Nachfolgend wird zwischen konvergenten und divergenten Leistungen unterschieden, um anschließend unterschiedliche Prüfungsformen hinsichtlich ihrer Vor- und Nachteile zu diskutieren.

Der schulische Leistungsbegriff wurde seit seiner Entstehung immer wieder kontrovers diskutiert. Mit der Entwicklung der Schule als Institution wurde die Frage, inwieweit von den Schülern und Schülerinnen Anforderungen erfüllt werden, also welche Leistung sie imstande sind zu erbringen, bedeutsamer. Furck (1975) arbeitet vier unterschiedliche Bedeutungen des Begriffs der Schulleistung heraus: Leistung bezogen auf den einzelnen Schüler kann als Anforderung an ihn von außen (von der Schule selbst) gesehen werden, als das von ihm erbrachte Ergebnis in spezifischen Leistungskontexten oder einfach als Beschreibung seiner schulischen Tätigkeit als solcher. Die vierte Bedeutung wäre schließlich nach Furck (1975) die Sicht auf Schulleistung als das, was die Schule *selbst* für die Gesellschaft, die Wirtschaft, die Wissenschaft etc. erbringt.

Die Kritik am Leistungsbegriff vor allem in den 1970er Jahren bezieht sich vor allem darauf, dass sie die Leistungsanforderungen an den Einzelnen „als repressives Verfahren einer einseitig auf Gewinnstreben ausgerichteten Herrschafts- und Wirtschaftsordnung“ (Vergl. Twellmann, 1981, Band VI, S. 330) versteht. Befürworter des Leistungsbegriffes dagegen betonen, dass die Entlohnung des Einzelnen in Abhängigkeit der von ihm erbrachten Leistung zu erfolgen hat, insofern eine Orientierung an der Leistungsfähigkeit notwendig ist. Twellmann (1981) argumentiert, dass je nach Position die Schule entweder als Abbild der Leistungsgesellschaft verstanden wird oder aber eine Stätte wird, an der Schüler und Schülerinnen die Leistungsverweigerung erlernen. Diesen Interpretationen entgegnend, spricht sich Twellmann (ebd.) für einen *pädagogischen Leistungsbegriff* aus, der zwar das Bemühen des Erfüllens der Anforderungen des Unterrichts einschließt, aber einen unerträglichen Leistungsdruck nicht aufkommen lässt.

Letzteres ist nicht nachvollziehbar, weil Twellmann damit impliziert, dass solange der Leistungsbegriff pädagogisch konnotiert ist, massiver Leistungsdruck ausgeschlossen ist. Betrachtet man aber beispielsweise aktuellere

Veröffentlichungen zur Leistungsmessung und -bewertung wie die von Paradies, Wester und Greving (2005), die ausschließlich von Lehrkräften und Pädagogen verfasst worden ist, findet man dort ein insgesamt doch deutlich an den Anforderungen des Arbeitsmarktes orientiertes Verständnis von Leistung.

„Am ehesten lassen sich Leistungsnormen in der Schule mit denen der Arbeitswelt verbinden. Insofern sozialisiert die Schule nicht nur mit Blick auf die Leistungen allgemein, sondern auch mit Blick auf die Anforderungen der Arbeitswelt. Das gelingt umso eher, als auch die so genannten Sekundär-Tugenden in unterschiedlichen Formen zum Tragen kommen. In Projekten und anderen kooperativen Arrangements lassen sich Sinn und Wirksamkeit erfahren. Deshalb ist es so wichtig, dass z.B. Projekte nicht von der Leistungsorientierung freigemacht bzw. von der Leistungsidee abgekoppelt werden." (Paradies, Wester & Greving, 2005; S.26).

Hier werden letztlich der Leistungsbegriff und die Forderung von Leistung nicht nur auf unterrichtliche Fähigkeiten, sondern darüber hinaus auch auf die persönliche Entwicklung bezogen und mit den Anforderungen der Arbeitswelt und der Gesellschaft gerechtfertigt.

Auch die von Ingenkamp und Lissmann (2008) vertretene Lösung, die Klärung des schulischen Leistungsbegriffs auf pragmatischer Ebene im Zuge einer „curricularen Einigung" (S. 132) zu erreichen, löst das in den 1970ern aufgeworfene Problem des Leistungsbegriffes nicht:

Gerade wenn Standards gesetzt werden, Lernziele sich nicht nur auf kognitive, sondern auch auf emotionale und soziale Lernbereiche erstrecken, ist dies ja gerade *kein* Beweis dafür, dass dies den Selektionsdruck verringert. Vielmehr trägt eine die Lernenden als Gesamtheit umfassende Beurteilung und Thematisierung im Rahmen von schulischer Leistungsbeurteilung, wie sie auch Paradies et al. (2005) vorschlagen, gerade dazu bei, dass der Freiheitsgrad der persönlichen Entwicklung durch eine derartige Standardisierung verringert wird, da die Ausrichtung der Individuen im Sinne einer Anpassung an vorgegebene Normen verstärkt wird.

Denn wer entscheidet, was Schüler und Schülerinnen zu leisten haben? Wie wird die Setzung von Standards diskutiert? Und was bedeutet eine Aussage wie „Die Pädagogik muss ein Mitspracherecht haben, um zu modifizieren und abzulehnen, was nicht der Entwicklung und der Mündigwerdung des Heranwachsenden dient." (Ingenkamp & Lissmann, 2008, S. 133)? Man könnte hier antworten, ob es nicht

sinnvoller wäre, den Heranwachsenden selbst ein Mitspracherecht einzuräumen, damit sie mündig werden können und sich zu selbstständigen Erwachsenen entwickeln können, anstatt durch *die Pädagogik* (wer immer auch für diese spricht) fremdbestimmt zu werden.

Auch die Erkenntnis von Paradies et al. (2005, S. 30), dass das von ihnen skizzierte Verständnis von Schulleistung womöglich wenig motivierend ist, führt nicht dazu, dass dieses in Frage gestellt wird:

„Die *Motivations- und Förderungsfunktion* ist ambivalent: Schüler werden durch Leistungsbeurteilung nicht nur motiviert, sich mit bestimmten Lerninhalten zu beschäftigen, sondern sie können so auch individuell gefördert werden. Andererseits muss eine sehr wichtige Einschränkung gemacht werden: da Leistungsbeurteilung immer auch eine selektive Wirkung hat, können Schüler dann, wenn sie nicht die gewünschte Leistung erbringen, sehr leicht demotiviert und blockiert werden! Wir sind daher der Meinung, dass es im Lernprozess immer auch beurteilungsfreie (Monitoring-) Phasen geben muss, und werden auf diesen Aspekt ausführlich später zu sprechen kommen."

Man könnte hier argumentieren, dass die formulierte Ausnahme („, dass es im Lernprozess immer auch beurteilungsfreie (Monitoring-)Phasen geben muss…" (S. 30)) eigentlich doch die Regel sein müsste. Wenn Schule ein Ort ist, an dem zunächst gelernt werden soll, dann muss ein Lernprozess, wenn er zum Erfolg führen will, durch Fehlertoleranz gekennzeichnet sein. Wie aber lässt sich Fehlertoleranz mit ständigem Monitoring und fortwährender Beurteilung vereinen? Werden Lernende fortwährend beurteilt, werden sie versuchen, möglichst keine Fehler zu machen, weil sie wissen, dass sich diese negativ auf ihre Beurteilung auswirken werden. Bei dem oben geschilderten Verständnis von Schulleistung tritt die Lernzielorientierung in den Hintergrund, die Leistungszielorientierung dagegen in den Vordergrund. Dass eine solche Orientierung langfristig schädigend für das Selbstkonzept ist, zeigen empirische Studien (Ståhlberg et al., 2019; Tuominen-Soini et al., 2008), hingegen eine Lernzielorientierung sich als förderlich für das selbstregulierte und selbstbestimmte Lernen erweist (vergl. Dweck & Leggett, 1988; Elliot & McGregor, 2001).

Insgesamt lässt sich feststellen, dass im Zuge der Formulierung von Bildungsstandards und Einziehung von Qualitätssicherungsmaßnahmen eine deutlich an vorgegebenen Normen orientierte und damit teilweise auch stark

verkürzte Sicht auf Leistungsbeurteilung erfolgt ist, die es erneut kontrovers zu diskutieren gilt.

Konvergente und divergente Leistungen

Im Rahmen der Leistungserfassung wird zwischen konvergenten und divergenten Leistungen unterschieden. Von konvergenten Leistungen sprechen wir dann, wenn diese ein eindeutiges Ergebnis haben und entsprechend gut messbar sind. Dies könnten korrekte Antworten auf Faktenfragen sein, Rechnen nach vorgegebenen Formeln etc. (Ingenkamp & Lissmann, 2015).

Divergente Leistungen dagegen sind schwer messbar, weil es verschiedene Ergebnisse geben kann, die dennoch gleichwertig sind. Die Einordnung der Leistung als falsch oder richtig ist entsprechend schwierig. Bei der Erfassung der divergenten Leistung ist es entsprechend bedeutsam, dass die zu erfassenden Leistungen klar definiert werden und entsprechende Erwartungshorizonte erarbeitet werden, die eine Einordnung der Qualität der Leistungen ermöglichen. Beispiele für divergente Leistungen gibt es viele: Angefangen von Deutschaufsätzen, Interpretationen (z.B. im Fach Geschichte), Kunstwerken, Kompositionen, aber auch originellen Lösungsansätzen für mathematische und naturwissenschaftliche Probleme fallen in diesen Bereich (vergl. Ingenkamp & Lissmann, 2008).

Mündliche Prüfungen

Unter einer mündlichen Prüfung versteht man eine Leistungserbringung einer oder mehrerer Kandidat*innen vor einem oder mehreren Prüfer*innen (Kommission). Die Fragen an die zu Prüfenden liegen entweder schriftlich vor oder werden mündlich gestellt. Die mündliche Prüfungsleistung wird entweder im Rahmen eines Prüfungsprotokolls registriert oder im Rahmen der nachfolgenden Beurteilung von den Prüfer*innen aus dem Gedächtnis rekonstruiert. Auch Mischformen sind denkbar.

Die Bedingungen einer mündlichen Prüfung können entweder durch eine Prüfungsordnung festgelegt werden (Dauer, Inhalt, zu erstellende Protokolle, Besetzung der Prüfungskommission etc.; formelle Prüfung) oder aber sie werden im Rahmen einer informellen Prüfung ohne festgelegte Regularien und unter Umständen ohne Protokollierung der Leistung durchgeführt.

Eine besondere Herausforderung stellt die Prüfung von mehreren Personen gleichzeitig dar (Gruppenprüfung). Allein schon aufgrund der zu gewährleistenden Rechtssicherheit muss eine Einzelbeurteilung der Kandidaten möglich und nachvollziehbar sein.

Beispiel

Mündliche Abiturprüfungen sind in der Regel formelle Prüfungen, da genau Umfang, Dauer und Zusammensetzung der Kommission festgelegt sind. Diese Prüfungen werden protokolliert und normalerweise liegen die Leitfragen schriftlich vor. Insgesamt ist die Standardisierung der Prüfung vergleichsweise hoch.

Sprachprüfungen in der Sekundarstufe I wären eine Mischform zwischen formeller und informeller Prüfung. Es liegen hier Rahmenbedingungen vor, die aufgrund von Absprachen in den jeweiligen Fachkonferenzen getroffen werden, die normalerweise als „Soll-Bestimmungen“ gehandhabt werden. Es gibt in der Regel ein Kurzprotokoll und einen groben Ablauf der Prüfung. Da in diesem Zusammenhang häufig auch Gruppenprüfungen erfolgen, gibt es mehr oder weniger ausgearbeitete Protokollbögen, die eine Einzelbeurteilung ermöglichen sollen.

Mündliche Überprüfungen des Gelernten im Rahmen der Beurteilung der Lernenden im Unterricht sind dagegen Beispiele für informelle Prüfungen. Dauer, Fragen etc. sind nicht festgelegt und die Verlaufsform ist von Zufällen oder Ad-hoc-Entscheidungen geprägt. Die Einschätzung erfolgt retrospektiv und wird auf Grundlage der von der Lehrkraft erinnerten Leistung vorgenommen.

Auch die Erfassung mündlicher Beteiligung zählt zu den informellen mündlichen Prüfungen. In Abhängigkeit der Erwartungen der Lehrkraft, kann mündliche Beteiligung per se im Sinne einer aktiven Teilnahme am Unterricht wertgeschätzt werden. Hier ist eine aktive Mitarbeit ausreichend, Fehler in den Antworten und Beiträgen der Schüler und Schülerinnen werden als Lerngelegenheit verstanden und werden entsprechend nicht negativ bewertet, sondern im Gegenteil konstruktiv gewendet. Diese Haltung ermutigt auch leistungsschwächere Schüler und Schülerinnen sich zu beteiligen, da sie keine negativen Konsequenzen für falsche Antworten befürchten müssen.

Beurteilt die Lehrkraft allerdings in ihrer Einschätzung der mündlichen Beteiligung (verbunden mit der entsprechenden Note am Ende eines Zeitraums) vor allem auch die *Qualität* der Beiträge, so ist davon auszugehen, dass Schüler und Schülerinnen abwägen, ob sie es riskieren können, einen Beitrag zu leisten. Das führt dazu, dass eine Lernkultur geschaffen wird, in der Lernende, die Defizite oder Verständnisschwierigkeiten haben, diese eher versuchen werden zu verdecken. Dieses wiederum verringert die Chance, dass Fehlvorstellungen von der Lehrkraft entdeckt werden. Sie wird bei einer höheren Anzahl korrekter Antworten auf das Verständnis der Klasse insgesamt schließen und auf diese Weise gerade schwächeren Schülern und Schülerinnen weniger (zusätzliche) Lerngelegenheiten anbieten. Entsprechend differenziert der Unterricht wenig und berücksichtigt in geringerem Maße die Heterogenität der Lernenden.

Inhalte mündlicher Prüfungen

Eine mündliche Prüfung bezieht sich naturgegeben auf sprachliche Leistungen bzw. auf an Sprache gebundene Leistungen. Diese können theoretisch durch eine Arbeitsprobe oder eine Veranschaulichung (Grafiken, Skizzen etc.) ergänzt werden. Ob dies der Fall ist, hängt von den jeweiligen Prüfungsmodalitäten fest, die die Prüfer*in festgesetzt haben.

Sprachliche Fertigkeiten, die Logik der Argumentation, aber auch die soziale und kommunikative Kompetenz, nämlich die Fähigkeit sich im Verlauf der Prüfung an die Vorstellungen und Erwartungen der Prüfer*in anzupassen, sind neben der inhaltlichen, fachlichen Darstellung des Wissens, zentrale Kriterien anhand derer Prüflinge in mündlichen Prüfungen beurteilt werden.

Charakteristika mündlicher Prüfungen

Im Rahmen von mündlichen Prüfungen besteht die Möglichkeit das Frageverhalten an die Leistung der Prüflinge anzupassen. Einerseits kann die direkte Interaktion zwischen Prüfer*innen und Prüflingen den Verlauf der Prüfung dadurch beeinflussen, dass sie durch verbale und nonverbale Reaktionen ermutigen (z.B. durch Lob, Kopfnicken, Lächeln etc.) oder verunsichern (z.B. durch Stirnrunzeln, mitleidsvollem Blick, negative Äußerungen etc.). Weiter haben die Prüfer*innen die Möglichkeit, in Abhängigkeit der Qualität der Beiträge anspruchsvollere Fragen zu stellen oder aber im Gegenteil die Schwierigkeit der Fragen abzusenken. In diesem Zusammenhang kann folgendes Problem

entstehen: Hat der Prüfende schon vor der Prüfung ein Urteil darüber gefällt, auf welchem Niveau die Fragen gestellt werden, kann es sein, dass ein Prüfling über- oder unterfordert wird.

Ein weiteres Charakteristikum mündlicher Prüfungen ist, dass sie häufig mit starker Angst besetzt ist (vergl. Moeller, 1972). Dieses hat Einfluss auf die Validität der Prüfung: Denn statt Leistung zu messen, wird unter Angstbedingungen eher die Fähigkeit gemessen, im Sinne einer Stressresistenz mit dieser umzugehen. Auch ist die Position naturgemäß hierarchisch. Die Prüflinge müssen sich an die vom Prüfer gewählte Prüfungsform anpassen. Das muss nicht, kann aber zu Willkür im Umgang mit Prüflingen führen. Diesem Problem widmen sich Prüfungsordnungen, die festlegen, wann ein Prüfling formal bestanden hat und dass beispielsweise Prüfer*innen keine sachfremden Erwägungen bei der Beurteilung des Prüflings anstellen dürfen.

Beispiel

Laura hat sich auf die mündliche Prüfung im Fach Biologie lange und intensiv vorbereitet, weil sie ihre bisherige Note (ausreichend) verbessern möchte. Als sie in die Prüfung geht, ist sie verwundert darüber, dass ihre Biologielehrerin nur Fragen stellt, die eigentlich das Basiswissen betreffen. Einerseits ist sie erfreut, dass sie alle Fragen leicht beantworten kann, andererseits aber fragt sie sich, warum sie sich so intensiv vorbereitet hat.

Als sie hereingerufen wird, teilt ihr die Biologielehrerin erfreut mit, dass sie eine „3“ bekommen hat, weil sie zeigen konnte, dass sie die Grundlagen verstanden hat. Laura beschwert sich und argumentiert, dass sie ja gar nicht zeigen konnte, was sie noch alles gelernt und verstanden hatte. Die Prüferin setzt dagegen, dass sie aufgrund der Vorzensur Laura nur auf eine „3“ geprüft habe, schließlich habe sie Laura nicht überfordern und frustrieren wollen.

In dem hier dargestellten Beispiel wird ersichtlich, dass die getroffene Vorannahme, dass die Schülerin keinesfalls mehr als „befriedigende Leistungen“ erreichen kann, dazu führt, dass sie egal wie gut sie gelernt hat, aus diesem Bewertungsrahmen nicht ausbrechen kann (allenfalls im Sinne einer deutlich schlechteren Leistung). Die Prüferin selbst unterstreicht ihre positive Absicht: Sie habe die Schülerin nicht demotivieren, sondern ihr die Gelegenheit geben wollen,

wenigstens Grundlagenwissen gut darstellen zu können. Die Prüferin begeht hier einen typischen Beobachtungsfehler: Ihre Objektivität ist beeinträchtigt, weil sie die Prüfung auf der Grundlage des Wissens um die Vorzensur der Schülerin gestaltet.

In diesem Zusammenhang stellt sich die Frage nach den Vor- und Nachteilen mündlicher Prüfungen.

Gerade frühe Untersuchungen zeigen, dass Probleme vor allem in der Gewährleistung der Einhaltung von Gütekriterien auftreten. Mündliche Prüfungen sind wenig objektiv, zuverlässig und gültig. Im Folgenden werden beispielhaft Studien zitiert, die zeigen, wo Fehlerquellen liegen können.

Tatsächlich zeigen Untersuchungen, dass in mündlichen Prüfungen sowohl Kontrasteffekte als auch das Wissen über die Vorzensur der Prüflinge eine Rolle bei der Einschätzung der Prüfungsleistungen spielen (vergl. Birkel, 1978). Hat ein schwächerer Schüler das Pech mit einem stärkeren Schüler zusammen oder direkt nach ihm geprüft zu werden, so wird seine Leistung signifikant schlechter wahrgenommen als wenn er nach einem in der Leistung vergleichbaren Schüler geprüft wird. Vorinformation beeinflussen ebenfalls die Einschätzung der Leistung (Birkel, 1978). Neuere Studien zeigen auch, dass Studierende mit ausländischem Namen im Fach Jura in mündlichen Prüfungen signifikant benachteiligt werden im Vergleich zu ihren Kommilitonen mit deutschem Namen (Towfigh, Traxler und Glöckner, 2014).

Auch die Redegeschwindigkeit hat einen Einfluss auf die Wahrnehmung der Qualität. Ein schnelleres Sprechtempo bei identischem Text wird mit einer durchschnittlich deutlich besseren Note bewertet (vergl. Birkel & Pritz, 1980).

Dennoch kann es auch gegensätzliche Effekte geben: Bietet die mündliche Prüfung den Prüflingen die Möglichkeit der Themenauswahl und der Steuerung der zu prüfenden Inhalte, so können sie dort auch erheblich besser abschneiden als in schriftlichen Arbeiten. Dies könnte zumindest universitäre Leistungsunterschiede in Abhängigkeit der Prüfungsform erklären (vergl. Grabowski, 1999).

Die Möglichkeit, die Fragen an die Voraussetzungen des Prüflings anzupassen, können dabei sowohl hilfreich sein als auch kontraproduktiv (vergl. Beispiel oben). Sicher ist, dass die Wiederholung im Sinne eines Retests einer mündlichen Prüfung nahezu ausgeschlossen ist.

Anpassungsleistungen des Prüflings an die Erwartungen der Prüfer*innen auch in Bezug auf Auftreten (Erster Eindruck, Erscheinung, Kleidung, Haltung) und Kommunikation spielen eine bedeutsame Rolle bei der Beurteilung (vergl. Jäger, 2007). Entsprechend kritisiert Lautmann bereits 1971, dass vor allem Unterschichtkinder in diesen Prüfungen benachteiligt werden. Damit wird aber das Gütekriterium der Validität verletzt.

Einige dieser Probleme lassen sich durch Standardisierung von Prüfungen und die Erstellung von Fragelisten und Erwartungshorizonten abmildern. So können Fragelisten erstellt werden, die auf unterschiedlichen Niveaustufen Leistung erfassen. Werden bei jedem Prüfling alle Niveaustufen angeboten, besteht immerhin eine Vergleichbarkeit bezogen auf den Schwierigkeitsgrad der Prüfung. Weiter können Protokolle erstellt werden, um Erinnerungsfehlern und einer eingeschränkten Sichtweise auf den Prüfling vorzubeugen. Kriterien, was eigentlich als „gut“ betrachtet wird, können vorab festgelegt werden, sodass Leistungsbeurteilungen transparent und nachvollziehbar gestaltet werden können.

Grundsätzlich sollte versucht werden, die Prüfungssituation immer so zu gestalten, dass möglichst keine Angst auftritt. Dies kann durch üben der Prüfungssituation gelingen, aber auch durch die Klärung von wichtigen Fragen vorab: Was mache ich als Prüfling zum Beispiel, wenn ich eine Frage nicht verstanden habe? Was mache ich, wenn ich einen Blackout habe?

Wie wir sehen, sind all diese Maßnahmen mit einem gewissen Aufwand verbunden und bedingen die Fähigkeit der Prüfenden, sich selbst und ihr Verhalten fortwährend kritisch zu reflektieren.

Mündliche Leistung – ein Sonderfall der mündlichen Prüfung

Vor diesem Hintergrund wird das Problem der mündlichen Leistungsbeurteilung offenbar: Die mündlichen Leistungen fließen zu einem in der Regel größeren Anteil als die schriftlichen Leistungen in die Gesamtnote in den Fächern ein. Das stellt die Lehrkraft aber vor die Aufgabe, diese sinnvoll und unter Einhaltung der Gütekriterien zu erfassen. Hier zeigt sich das bereits an verschiedenen Stellen beschriebene Dilemma:

Wann und wie wird mündliche Leistung überhaupt erfasst? Im Rahmen von festgelegten Prüfungsabschnitten wäre eine Kontrolle von unerwünschten

Beobachtungsfehlern und Einflüssen vielleicht noch machbar. Als Einschätzung einer irgendwie gesamt erinnerten und global eingeschätzten mündlichen Leistung wird die Erfassung durch eine Vielzahl von Fehlern alles andere als objektiv, reliabel oder valide sein. Weiter ist die Frage, was ich als Kriterium einer gelungenen mündlichen Beteiligung festlege? Ist es die *Lernbereitschaft* der Schüler und Schülerinnen? Dann wäre die auf den Unterricht gerichtete Aktivität zu erfassen, die sich nicht notwendigerweise in tatsächlichen Redeanteilen zeigt, sondern auch in der Aufmerksamkeit oder Fokussierung auf den Unterrichtsgegenstand. Auch wäre es irrelevant, ob und wie oft Fehler gemacht werden, da es um die Lernbereitschaft nicht um den Lernerfolg ginge. Letzterer würde eher im Rahmen einer schriftlichen Prüfung erfasst.

Oder aber will ich die mündliche Beteiligung im Sinne einer mündlichen Leistung erfassen und Beiträge hinsichtlich ihrer Qualität einschätzen. Dann wäre die Frage, ob dies fortlaufend (also immer im Unterricht) geschieht oder nur zu ausgewählten Zeitpunkten (Referate, mündliche Prüfungen).

Um nur einige Probleme zu nennen, die bei der globalen Einschätzung der mündlichen Aktivität von Schülern und Schülerinnen auftreten, werden im Folgenden einige kritische Fragen benannt:

Wo sitzen die Schüler und Schülerinnen? Nimmt die Lehrkraft deren Engagement überhaupt gleichermaßen wahr? Wann wird eine „falsche Antwort“ als zulässig im Sinne des voranschreitenden Lernprozesses negativ oder positiv bewertet? Wie beeinflusst die Lehrkraft die Art und Weise der Beiträge? Welches Gewicht misst sie „unangemessenen“ Beiträgen (Witze, Störungen, Provokationen) bei und wie verändern diese die Gesamteinschätzung der mündlichen Leistung?

Angesichts dieser Fragen sehen wir, dass wiederum die Grenze zwischen Anpassungsanforderungen (Betragen etc.) und tatsächlicher Leistung verschwimmt, die Einschätzung der mündlichen Leistung wird auch an Wohlverhalten gekoppelt, also ein Instrument der Disziplinierung. So könnte ein Schüler, der sich beispielsweise aktiv und mit guten Beiträgen beteiligt, weil er auch mal eine provokante Bemerkung macht, mit einer schlechteren Note abgestraft werden. Tatsächlich beeinflusst und verzerrt die Wahrnehmung und Einschätzung des Sozialverhaltens von Schülern und Schülerinnen die Beurteilung ihrer fachlichen Leistungen (Krämer & Zimmermann, 2020): So zeigt sich, dass die tatsächlichen Fachleistungen (mündlich und schriftlich) etwa zu 53 % die Fachnote bestimmen, das Unterrichts*verhalten* dagegen zu 31 % und zu weiteren 16 %

Disziplin- und Sozialverhalten die Fachnote bestimmen (Fiske & Neuberg, 1990; Heyder, Kessels & Retelsdorf, 2019).

Schriftliche Prüfungen

Grundsätzlich muss ausgehend vom Ziel der Untersuchung zwischen informellen Verfahren, die domänenspezifische Lernergebnisse im Rahmen von Klausuren und Tests erfassen wollen, und bildungswissenschaftlichen Verfahren, deren Ziel die Feststellung domänenspezifische Kompetenzausprägungen zum Zwecke der Forschung, des Vergleichs von Bildungssystemen und deren Monitoring unterschieden werden (vergl. Spinath & Brünken, 2016). Darüber hinaus stellen auch psychologische und pädagogisch-psychologische Testverfahren schriftliche, standardisierte Prüfungen dar.

In diesem Kapitel beschäftigen wir uns zunächst ausschließlich mit den informellen Verfahren (Tests/Klausuren etc.) im Rahmen der schulischen Verlaufs- und Ergebnisdiagnostik durch Lehrkräfte.

Unter schriftlichen Prüfungen versteht man alle Prüfungsformen, bei denen Prüflinge Aufgaben, die ihnen mündlich oder schriftlich gestellt werden, schriftlich bearbeiten (vergl. Jäger, 2007). In der Regel liegen zudem Anweisungen vor, die den Prüflingen darüber Auskunft erteilen, ob es formelle Vorgaben (Korrekturränder, zur Verfügung stehende Zeit, Hinweise zum Umgang mit Täuschungsversuchen) gibt und möglicherweise, wie die Aufgaben bearbeitet werden sollen (z.B. Anfertigungen von Skizzen, das Gebot der Darlegung von Lösungswegen, auf Papier oder digital).

Lerntagebücher und Portfolios werden gesondert betrachtet, da sie anders aufgebaut sind und auch eine andere Zielsetzung verfolgen.

Schriftliche Prüfungsformen sind aus Schule, Ausbildungsstätte und Universität nicht wegzudenken. Sie stellen das gängige Verfahren dar, um Leistungen zu überprüfen und Standards zu sichern. Gemeinhin verknüpft man mit der schriftlichen Prüfung die Annahme, dass sie objektiver ist als die mündliche Prüfung, da Urteile mehrfach und auch von unabhängigen Prüfer*innen jederzeit überprüft werden können. Zudem lässt sich nur in einer schriftlichen Prüfung eine gewünschte Anonymisierung erreichen. Letzteres ist in der Realität der Erfassung von Schulleistungen mit schriftlichen Prüfungen in der Regel nicht gegeben, da die

mit Namen genannten Prüflinge bekannt sind (eine Ausnahme bilden Abiturprüfungen, wo die Zweitprüfer unter Umständen die Prüflinge nicht kennen). Zudem kann auch das bloße Nennen eines Namens die Auswertungsobjektivität einschränken, doch dazu später mehr.

Schriftlichen Prüfungsformen werden dabei eine Reihe von Funktionen zugeschrieben. Da die schriftliche Prüfung als objektiv angesehen wird, wird sie zur Selektion und Zuweisung zu bestimmten Tätigkeitsfeldern oder Bildungschancen genutzt. Dadurch, dass schriftliche Prüfungen auch von den Prüflingen wiederholt betrachtet werden können, kann die inhaltliche Rückmeldung über Stärken und Schwächen durch die Lehrkraft zielgerichteter und nachvollziehbarer erfolgen. Letzteres kann zu einer verbesserten Transparenz der Bewertung führen. Was wiederum auch die Möglichkeit beinhaltet, im Zweifelsfall die Bewertung gerichtlich anzufechten.

Auch schriftliche Prüfungen können differenziert angeboten werden (z.B. bezogen auf den Schwierigkeitsgrad oder den Umfang). Anders als mündliche Prüfungen sind sie aber, da mehrere Prüflinge die gleichen Aufgaben erhalten, auch besser interindividuell vergleichbar.

Typische Beispiele für schriftliche Prüfungen sind:

- Aufsätze, die Bildbeschreibungen, Text- und Bildinterpretationen als auch Erlebnisse thematisieren können.
- Schriftliche Arbeiten in Form von verschriftlichen Referaten, Essays zu bestimmten Themen oder Hausarbeiten.
- Diktate in eigener oder fremder Sprache, als Fließtext oder in Form von Lückentexten.
- Problemlöseaufgaben: Hier werden konkrete Aufgaben vorgegeben, die zu bearbeiten bzw. zu beantworten sind. Letztlich können diese in (fast) jedem Fach erstellt werden.
- Vokabeltests
- Multiple-Choice-Aufgaben oder Ja/Nein-Statements; hier muss aus einer oder mehreren Antwortalternativen die korrekte ausgewählt werden.
- Lückentexte, bei denen Fachbegriffe bzw. bestimmte Wörter fehlen, die entweder frei ergänzt werden müssen oder aus einer Liste zugeordnet werden müssen.
- Problemstellungen, welche theoretisch in allen Unterrichtsfächern konzipiert werden können. Mathematische Aufgaben, Analysen von

Texten, Quellen, Bildern etc. sind ebenso denkbar wie aufeinander aufbauende Problemstellungen (z.B. in den Natur- oder Gesellschaftswissenschaften)

Aufsätze und offene Problemstellungen

Aufsätze, schriftliche Arbeiten mit Problemstellungen sowie schriftliche Arbeiten zu bestimmten Themenstellungen (Referate, Hausarbeiten etc.) haben zunächst vergleichbare Probleme hinsichtlich der Einhaltung der Gütekriterien Objektivität, Reliabilität und Validität.

Aufgaben, die nicht eindeutig lösbar sind, also eine Vielfalt von Lösungswegen denkbar erscheinen lassen, bergen das Problem, dass unklar ist, wann eine Lösung besser oder schlechter, ausreichend oder ungenügend ist. Hier müssen Lehrkräfte zunächst das, was sie prüfen wollen operationalisieren, d.h. sie müssen festlegen, woran sie denn genau erkennen können, ob und in welchem Umfang eine Kompetenz erworben wurde und in welchem Verhalten sie sich zeigt. Gerade in aktuellen Curricula stellt dies ein größeres Problem dar, da die genannten Fähigkeiten häufig in Kompetenzclustern zusammengefasst sind und derart komplex sind, dass sie in der Regel nicht im Rahmen einer Arbeit erfasst werden können. Entsprechend müssen Lehrkräfte Teilkompetenzen soweit herunterbrechen, dass sie für sie erkennbar und damit auch prüfbar werden.

Zahlreiche Studien liegen zu Problemen der Auswertungs- und Interpretationsobjektivität vor. Gibt man den Prüfer*innen zusätzliche Informationen über die soziale Herkunft oder werden Namen vergeben, die auf einen Migrationshintergrund oder/und einen niedrigeren sozialen Status hin gedeutet werden können, so zeigt sich, dass diese Merkmale die Leistungseinschätzung beeinflussen. Besonders im Bereich der Rechtschreibung werden gleiche Leistungen aufgrund divergierender Informationen über die Personen sehr unterschiedlich eingeschätzt. Dieser Effekt zeigt sich auch in den Bereichen Stil und Inhalt (vergl. Ingenkamp und Lissmann, 2008). Geschlechterstereotype und die Bildung von Stereotypen aufgrund von außerunterrichtlichen Informationen (Einschätzung der Person durch andere, Interessen etc.) gelten ebenfalls als Ursache für Urteilsverzerrungen (vergl. Heyder et al., 2019).

Auch bezogen auf die Urteilsübereinstimmung zeigt sich, dass Lehrkräfte gleiche Leistungen (z.B. Aufsätze; Birkel & Birkel, 2002) sehr unterschiedlich einschätzen

und dies auch in Fächern mit eindeutig lösbaren Aufgaben wie in Mathematik (Birkel, 2005). Erwartungshorizonte können dieses Problem verringern, aber nicht beseitigen, da auch diese wiederum unterschiedlich interpretiert werden können.

Darüber hinaus ergeben sich Probleme, wenn den Schülern und Schülerinnen der Erwartungshorizont nicht bekannt ist. Zusätzlich zu den oben genannten Beurteilungsfehlern tritt nämlich hier das Problem auf, dass die Lernenden selbst die Aufgabenstellung interpretieren müssen. Sie müssen also antizipieren, was die Lehrkraft von ihnen erwartet. Je nachdem wie gut ihnen das gelingt, können sie mehr oder weniger zeigen, was sie gelernt und verstanden haben.

Schließlich ist auch die Stabilität von Lehrerurteilen problematisch, wenn sich die Leistungen der Schüler und Schülerinnen verändern. Untersuchungen zeigen, dass Leistungseinschätzungen zu wenig an verändertes Verhalten bzw. veränderter Leistung von Schüler und Schülerinnen angepasst werden. Sie sind de facto stabiler als die Merkmale, die sie erfassen wollen.

Multiple-Choice-Tests

Multiple-Choice-Tests und Lückentexte (bei vorgegebenen Alternativen) haben den Vorteil, dass sie sich objektiv auswerten lassen, dennoch stellen sich hier andere Probleme: Sind die alternativen Antworten wirklich eindeutig und unmissverständlich formuliert? Sind sie insofern trennscharf, als dass sie hilfreich sind, um eindeutig zu identifizieren, ob ein Schüler oder eine Schülerin tatsächlich den Stoff verstanden hat oder nicht? Oder anders ausgedrückt: Kann der Test zuverlässig herausfinden, wer welchen Lernstand erreicht hat? Hier liegen die Probleme vorgelagert in der Konstruktion des Testes selbst. Zudem muss darauf geachtet werden, dass das Testverfahren die abzufragenden Lernziele sinnvoll abbildet.

Soll es gelingen die Leistung der Schüler und Schülerinnen zuverlässig zu beurteilen, dann ist bei der Konzeption darauf zu achten, dass die Aufgabenschwierigkeit eher im mittleren Bereich liegt und nur wenige Aufgaben besonders leicht oder schwer zu lösen sind. Inwiefern dies gelungen ist, kann allerdings nur eine nachfolgende statistische Itemanalyse klären. Sinnvollerweise werden deshalb MC-Tests im Sinne eines Aufgabenpools über viele Jahre hin weiterentwickelt und immer wieder überprüft.

Portfolios und Lerntagebücher

Prinzipiell stellen sich bei der Beurteilung von Portfolios und Lerntagebüchern ähnliche Probleme wie bei Aufsätzen und Essays. Man kann darüber streiten, ob diese überhaupt Gegenstand der Beurteilung werden sollten oder ob sie im Sinne einer Prozessdiagnostik genutzt werden sollen, um Förderbedarfe zu entdecken und entsprechend den Unterricht zu modifizieren und Schüler und Schülerinnen individuell zu fördern.

Beispiel

Im Rahmen einer Lektüre in der neunten Klasse bearbeiten alle Schüler und Schülerinnen eine Lesemappe. Hier gibt es Pflichtaufgaben und Wahlaufgaben.

Der Erwartungshorizont der jeweils gestellten Aufgabe ist genau definiert.

Marjella, deren Eltern beide als Juristen arbeiten, hat aus den Wahlaufgaben der Lesemappe die Aufgaben ausgewählt, die stärker mit dem Text des Buches arbeiten und weniger kreative Freiräume eröffnen (wie z.B. das Erstellen eines Titelbildes oder das Schreiben eines alternativen Endes des Buches).

Als Marjella die am Ende eingereichte Mappe zurückbekommt, hat sie eine „3“ erhalten, obwohl sie aus ihrer Sicht alle Aufgaben zuverlässig anhand des Erwartungshorizonts abgearbeitet hat.

Sie fragt ihre Lehrkraft, warum sie am Ende nur eine „3“ bekommen hat. Diese sagt: „Tja, Marjella, von dir hätte ich schon mehr erwartet. Schließlich hast du ja zuhause Hilfe und hättest auch bei der Wahl der Aufgaben zeigen können, dass du bereit bist dich anzustrengen!“

In dem Moment, wo sie Gegenstand einer Leistungsbeurteilung werden, stellt sich zudem das Problem, dass hier häufig nicht nur die Leistung selbst, sondern auch die Leistungsbereitschaft (Motivation), persönliche Interessen und Einstellungen und soziale Herkunft (in Form von Unterstützungsangeboten in der Familie) der Schüler und Schülerinnen zum Gegenstand der Beurteilung werden. Zumindest können all diese Faktoren die Qualität des Portfolios selbst beeinflussen und

entweder dadurch oder durch das Wissen der Lehrkraft um diese Faktoren auch den Prozess der Urteilsfindung beeinflussen.

Im Beispiel oben, treten verschiedene Beurteilungsfehler auf, weil die Lehrkraft davon ausgeht, dass Marjella auf jeden Fall bessere Voraussetzungen als andere hatte und sich das im Portfolio hätte niederschlagen müssen. Es ist aber gar nicht gesagt, ob und wie Marjella durch die wissenschaftliche Tätigkeit ihrer Eltern profitiert hat. Weiter zeigt die Lehrkraft nachträglich durch ihre Aussage, dass die Wahlaufgaben eigentlich gar keine Wahlaufgaben im eigentlichen Sinne waren, sondern die Wahl selbst schon einen Einfluss auf die Beurteilung hatte. Die Lehrkraft hat daraus Rückschlüsse auf die Motivation der Schülerin und ihr Engagement gezogen. Auch macht sie deutlich, dass sie sich bereits *vorher* eine Erwartung gebildet hatte, die Marjella hätte erfüllen sollen.

Schließlich sollte die Frage des Interesses oder der Motivation an der Bearbeitung einer schulischen Aufgabe nicht die Beurteilung der gezeigten Leistung beeinflussen.

RÜCKMELDUNG VON LEISTUNGSBEWERTUNGEN

Die Rückmeldung von Leistungsbewertungen, von Lernentwicklungsdiagnosen oder auch von sozialem Verhalten erfolgt im schulischen Bereich in unterschiedlichster Form: Einerseits stellt das *Ziffernzeugnis* bzw. die *Zensur* (bei schriftlichen Arbeiten, nach mündlichen Prüfungen etc.) eine in einer Note verdichtete Rückmeldung dar, andererseits können die Bewertungen in Form von Ziffern auch von *mündlichen oder schriftlichen Ausführungen* begleitet sein. Auch eine Rückmeldung, die gänzlich auf Noten verzichtet z.B. im Rahmen eines *Lernentwicklungsberichts* ist denkbar.

Im Folgenden werden wir uns zunächst mit den grundsätzlichen Zielen von Rückmeldungen befassen. Im Anschluss daran werden unterschiedliche Formen der Leistungs- und Entwicklungsrückmeldung thematisiert und deren Vor- und Nachteile herausgearbeitet.

In der Bremischen Zeugnisverordnung wird ausgeführt, dass Zeugnisse und Lernentwicklungsberichte eine zusammenfassende Beurteilung der Lernentwicklung des Schülers oder der Schülerin in einem bestimmten Zeitabschnitt geben und der Unterrichtung sowohl der Schüler und Schülerinnen als auch deren Erziehungsberechtigten dienen. Darüber hinaus haben sie eine informative Funktion bei Übergängen auf andere Schulen als auch eine selektive Funktion bei Auswahlverfahren im Rahmen von Bewerbungen (Beruf und Ausbildung). Zudem sind sie die Grundlage, um Versetzungsentscheidungen zu treffen und Zugänge zu ermöglichen oder zu verwehren (z.B. die Möglichkeit das Abitur abzulegen).

Auch im laufenden Schuljahr erfolgen Rückmeldungen mit dem Ziel über die Leistungsentwicklung und den Lernstand der Schüler und Schülerinnen zu informieren. Hier liegt allerdings der Fokus auf der Optimierung von Lern- und Entwicklungsprozessen.

Grundsätzlich sollte eine Leistungsrückmeldung das Verhalten beschreiben und keine persönlichen Zuschreibungen („Du bist klug“, „Das liegt dir nicht“) enthalten. Sie sollte wertschätzend sein und möglichst Perspektiven aufzeigen.

Während eine kritische Rückmeldung an *Schüler und Schülerinnen* häufig dazu dienen soll, diese zu mehr Leistung oder Einsatz zu motivieren, kann eine Rückmeldung an die *Eltern* Anlass sein, diese zu beraten und gemeinsam

auszuloten, wie das Kind in seiner Weiterentwicklung unterstützt werden kann. Kritisch ist es, wenn Rückmeldungen allein im Sinne einer Aufforderung an die Lernenden oder die Erziehungsberechtigten gesehen werden, ohne dass die Lehrkraft ihr eigenes Handeln, ihren Unterricht und ihre Möglichkeiten den Entwicklungsprozess zu fördern mit reflektiert. Dann nämlich wird die Verantwortung für den Entwicklungsprozess einzig und allein an die Lernenden und womöglich noch die Eltern abgegeben. Letztlich muss aber beachtet werden, dass die Informationen, die sich aus der Bewertung einer Leistung, eines Verhaltens oder einer Kompetenz ergeben, im schulischen Bereich dazu dienen sollen, Maßnahmen zu ergreifen, um die Lernenden in ihrer Entwicklung zu *unterstützen*. Dies schließt eine kritische Reflexion des eigenen Handelns mit ein.

In ihren Publikationen weisen bereits die meisten Landesinstitute für Bildung darauf hin, dass eine Rückmeldung differenziert und informativ sein soll und dass sie auf der Grundlage transparenter Beurteilungsmaßstäbe getroffen werden soll. Weiter muss sie Verhalten beschreiben ohne die Person als solche abzuwerten. Man unterscheidet summative und formative Leistungsrückmeldungen.

Summative Leistungsrückmeldungen

Summative Leistungserhebungen dienen dazu, den individuellen Lernstand zu einem bestimmten Zeitpunkt (häufig z.B. am Ende einer Lernphase) festzustellen und zu bewerten. Neben schriftlichen Arbeiten, sind hier auch Präsentationen, Tests, Portfolios oder mündliche Prüfungen denkbar.

Summative Leistungsrückmeldungen beeinflussen künftige Lernprozesse insofern, als dass beispielsweise bei mangelnder Kompetenz nachgearbeitet werden muss. Ansonsten besteht die Gefahr, dass daran anschließende Lernprozesse beeinträchtigt werden (ein Beispiel dafür wäre fehlendes Grundwissen im Bereich „englische Grammatik“, das auch nachfolgend den weiteren Aufbau von grammatikalischem Wissen erschwert).

Die Rückmeldung über die festgestellten Kompetenzen erfolgt häufig über Angaben von Prozenträngen, Punkten oder Noten. Bedeutsam ist in diesem Zusammenhang auch welche Bezugsnormorientierung von der Lehrkraft bei der Beurteilung verwendet wird.

Formative Leistungsrückmeldungen

Bewertungen, die *während* des Lernprozesses vorgenommen werden, bezeichnet man als formative Leistungsbewertungen. Sie sollen dazu dienen, Rückmeldungen darüber zu geben, was schon erfolgreich erlernt wurde und wo andererseits noch

Fähigkeiten ausgebaut und Lernstände erworben werden müssen. Sie dienen nicht nur den Schülern und Schülerinnen, sondern auch den Eltern und Lehrkräften als Information, um den individuellen Lernprozess zu optimieren. Im Rahmen von formativen Rückmeldungen können unterschiedliche Perspektiven genutzt werden. Sie können aus der Perspektive der Lehrkraft, der Mitschüler*innen/Lernpartner*innen ebenso wie aus der Perspektive der Lernenden im Sinne einer Selbsteinschätzung vorgenommen werden. Formative Rückmeldungen können im Rahmen von Gesprächen, mit Hilfe von Lerntagebüchern oder Kompetenzrastern erfolgen. Auch Selbst- und Fremdeinschätzungsbögen können Grundlage einer Rückmeldung sein, z.B. im Rahmen eines Vortrags. Zu berücksichtigen ist auch hier, welche Bezugsnormorientierung bei der Rückmeldung genutzt wird. Gerade im Rahmen der formativen Rückmeldung sind die individuelle und die sachliche Bezugsnorm wichtig, da sie Schülern und Schülerinnen Hinweise über die eigene Entwicklung und ihren derzeitigen Entwicklungsstand geben können.

Auch aus Klassenarbeiten können Erkenntnisse für den weiteren Lernprozess gewonnen werden, wenn die Bearbeitung einzelner Aufgabenbereiche in den Blick genommen wird. Ebenso können aber auch umgekehrt Ergebnisse aus Lernnachweisen zur Kompetenzerreichung mit einer Note beurteilt werden. Formative Leistungsrückmeldungen münden tatsächlich häufig in eine summative Leistungsrückmeldung. Dies wird am Beispiel der Beurteilung mündlicher Leistung deutlich: Hier werden formative Leistungsrückmeldungen in der Regel aufsummiert und dienen als Bewertungsgrundlage für die abschließende Zensur. Dennoch ist zu unterscheiden zwischen einer Leistungsrückmeldung mit dem Ziel der Förderung der Lernentwicklung und einer Leistungsbewertung im Sinne eines Urteils (vergl. dazu Leistung diagnostizieren). In diesem Zusammenhang stellt sich erneut die Frage, ob die Schule ein Raum sein soll, in dem Kompetenzen erkannt und gefördert, Fehler reflektiert und dadurch bearbeitbar gemacht werden sollen, oder ob sie primär ein Ort ist an dem Wissen erworben und abgeprüft werden und die Verantwortung für den Wissens- und Kompetenzerwerb in erster Linie den Schülern und Schülerinnen zugeschrieben werden soll.

Numerische Beurteilung

Die Beurteilung mit Hilfe von Noten oder Punkten, die wiederum in Noten umgerechnet werden können, hat eine lange Tradition. Dabei ist die Kritik am Notensystem umfangreich: Noten sind zunächst wenig informativ. Sie stellen in

der Regel eine Verdichtung von Eindrücken und Bewertungen dar. Es ist aus ihnen nicht zu ersehen, welche Kompetenzen erworben wurden oder wo es Probleme gab. So kann eine „3“ in Deutsch bedeuten, dass das Kind in der Rechtschreibung sehr sicher (sehr gut) ist, aber im Ausdruck (ausreichend) und in der Produktion von Texten (ausreichend) größere Probleme hat. In diesem fiktiven Beispiel käme bei einer Gleichgewichtung der drei Bereiche am Ende eine „3 “ als Note heraus, ohne dass das Kind und die Eltern die zugrundeliegenden Informationen erkennen könnten. Aus der Note selbst ist zudem nicht ersichtlich, welche Bezugsnorm ihr zugrunde liegt oder wie zugrundliegende Einzelwertungen gewichtet wurden.

Ein weiterer Kritikpunkt ist die Tatsache, dass Leistungen in standardisierten Tests in nur geringem Maße mit Schulnoten korrelieren, das liegt auch daran, dass in Noten noch andere Bewertungen über den Schüler oder die Schülerin (Verhalten, Motivation, soziale Anpassung) eingehen, ohne dass dieses wirklich explizit erkennbar ist. Noten wirken zunächst objektiv und sachlich, auch wenn sie teilweise mit großen Messfehlern behaftet sind (Krämer & Zimmermann, 2020, Kaiser et al., 2013 Kaiser et al., 2015, Kaiser et al., 2017).

Befürworter von Ziffernzensuren betonen dagegen die Einfachheit und Plausibilität der Beurteilung und weisen darauf hin, dass sich Schüler und Schülerinnen miteinander vergleichen wollen und sowohl Schüler und Schülerinnen als auch deren Eltern die eindeutige Rückmeldung über Ziffern besser einordnen könnten als lange Berichte, die sie abschließend interpretieren müssten.
Eine alleinige Leistungsrückmeldung über Ziffern bleibt dennoch wenig informativ und hilft den Lernenden nicht weiter. Auch welche Bezugsnorm der Beurteilung mit Ziffern zugrunde gelegen hat, ist für Außenstehende nicht erkennbar!

Sie kann aber mündlich oder schriftlich erläutert werden, um den Informationswert zu erhöhen und auch zu klären, wie sie sich genau zusammensetzt und rechtfertigen lässt.

Rückmeldung über Kompetenzstufen

Eine Mischform zwischen Berichtszeugnis und Ziffernzeugnis ist die Rückmeldung über Kompetenzstufen. Ein Vorteil ist, dass sich hier die Kompetenzen breit auffächern lassen und so der Informationsgehalt in der Regel deutlich höher ist als beim reinen Ziffernzeugnis. Lehrkräfte, die eine Klasse übernehmen, Schüler und Schülerinnen und deren Eltern können erkennen, welche Kompetenzen in welchem Umfang erworben wurden. Je differenzierter die Kompetenzbereiche

beschrieben sind, umso höher der Informationsgehalt. Auch erfolgt hier eher eine Ausrichtung an einer sachlichen Bezugsnorm, da diese durch die Vorgabe bereits gegeben und intendiert ist. Dennoch kann man auch hier nicht erkennen, ob das Urteil, dass beispielsweise dazu führt, dass ein Kind in einem Bereich mit „hat die Kompetenz erreicht“ beurteilt wird, aus einem z.B. sozialen Vergleich mit der (vielleicht leistungsschwachen) Klasse herrührt. Die Einschätzung auf Kompetenzskalen kann in der Regel relativ leicht in Zensuren überführt werden, dies ist unter anderem ein Grund, warum sie eingesetzt werden. Wechseln beispielsweise Schüler und Schülerinnen von einer Schule, die Kompetenzraster in den Lernentwicklungsberichten nutzt auf eine Schule, die Notenzeugnisse verwendet, kann relativ einfach eine Umrechnung in Noten erfolgen.

Kompetenzstufen in Zeugnissen oder anderen Leistungsrückmeldungen sollten möglichst so formuliert werden, dass sie von den Schülern und Schülerinnen und deren Eltern auch verstanden werden. Eine Möglichkeit dies zu gewährleisten ist die mündliche Nachbesprechung im Rahmen von Elternsprechtagen nach der Zeugnisvergabe.

Die Zuverlässigkeit der Einschätzung der Schüler und Schülerinnen hängt in diesem Fall nicht unerheblich davon ab, ob sich die Lehrkräfte einig darüber sind, was in einem bestimmten Kompetenzbereich genau erwartet wird. Dies zu klären, ist beispielsweise eine Aufgabe von Fachkonferenzen.

Berichtszeugnisse und Lernentwicklungsberichte

Die Verwendung von Lernentwicklungsberichten und anderen Berichtsteilen in Zeugnissen hat in den letzten Jahren vor allem vor dem Hintergrund der Einführung von Oberschulen und Integrierten Gesamtschulen zugenommen. Die Intention ist es, Leistungen differenziert zurückzumelden und gleichzeitig in heterogenen Lerngruppen Vergleiche untereinander, die zu erhöhtem Konkurrenzdruck führen, zu verhindern.

Ziel des Lernentwicklungsberichtes ist es, eine differenzierende Beurteilung über den individuellen Entwicklungs- und Leistungsstand des Kindes oder Jugendlichen zu geben. In der Regel gibt es dazu Vorlagen vom Kultusministerium, die als verbindliche Orientierung dienen.

Ein Berichtsteil zum Lern- und Sozialverhalten ist nicht nur im Zusammenhang mit Lernentwicklungsberichten bekannt. Dieser wird häufig schon in den

Grundschulen verwendet. Allerdings zuweilen in einer Art, die dem beschriebenen Ziel zuwiderläuft: Wenn dort z.B. lediglich steht, dass das Arbeitsverhalten den Erwartungen entspricht, dann wird damit eigentlich ausgesagt, dass der Schüler eine „3“ erhalten hat. Hier wird die Note sozusagen standardisiert umschrieben und der Informationsgehalt ist entsprechend gering, Wir können nicht ablesen, was denn tatsächlich beurteilt wurde. Was zum Beispiel ist Gegenstand der Beurteilung im Bereich “Sozialverhalten” gewesen? War es die Anpassungsfähigkeit an das schulische System? War es das soziale Verhalten im Umgang mit Gleichaltrigen? Gleiches gilt für die Kurzbeschreibung des Lern- und Sozialverhaltens im Rahmen von Ziffernzeugnissen an Gymnasien.

Schulen, die tatsächlich zumeist in den Jahrgängen 5 bis 7 ausschließlich Lernentwicklungsberichte nutzen, unterteilen diese in einen allgemeinen Bericht zum Lern-, Arbeits- und Sozialverhalten und einen fachlichen Teil, der daraus besteht, dass zu jedem einzelnen Fach eine verbale Rückmeldung gegeben wird. Dies kann durch Kompetenzprofile ergänzt werden.

Die Anforderung an einen Lernentwicklungsbericht ist, dass sich dieser sowohl auf den individuellen Entwicklungsstand bezieht als auch auf den Leistungsstand, gemessen an der kriterialen Bezugsnorm (also z.B. an erreichten Kompetenzen). Er darf keine Formulierungen enthalten, die eine verdeckte Benotung darstellen, weil dies der Intention des Lernentwicklungsberichtes zuwiderläuft.

Kritiker der Berichtszeugnisse verweisen darauf, dass die Erstellung dieser sehr aufwändig sei, weil jeder Schüler und jede Schülerin individuell betrachtet werden müsse. Dem ist zu entgegnen, dass jede vernünftige Diagnostik und Leistungserfassungen und -rückmeldungen aufwändig sind. Wer gut begründet Noten verteilen möchte oder muss, muss im Entscheidungsprozess eine vergleichbar aufwändige Diagnostik betreiben.

Dennoch gibt es auch Kritik, die plausibel erscheint: So kann eine Tendenz beobachtet werden, Textbausteine zu verwenden oder sich an vorigen Zeugnissen zu orientieren und dadurch unerwünschte Beurteilungsfehler zu begehen. Kritiker mahnen auch, dass ein Berichtszeugnis zwar scheinbar auf der Grundlage einer eingehenden Diagnostik erfolgt, dies kann aber keinesfalls vorausgesetzt werden. Die Studie von Jürgens (2001) zeigt, dass Lehrkräfte häufig ihre Urteile nicht auf systematische Beurteilungen stützen, sondern vielmehr auf einen Gesamteindruck, der sich aus verschiedenen Quellen

(Gelegenheitsbeobachtungen, Unterrichtsbeobachtungen) speist. Letztlich ist der Bericht also nur so gut wie die vorangegangene Diagnostik.

Ein weiterer Kritikpunkt ist der, dass durch die Fokussierung auf über- und außerfachliche Kompetenzen wie Arbeits- und Sozialverhalten Schüler und Schülerinnen als Gesamtheit in den Blick geraten und somit als Person beurteilt werden. Diese Kritik steht nur scheinbar im Widerspruch zu der pädagogischen Forderung nach einer ganzheitlichen Betrachtung von Schülern und Schülerinnen. Das Problem besteht weniger in der Beschreibung von Verhalten als vielmehr darin, dass damit auch häufig eine Normierung einhergeht: Die Lehrkraft entscheidet im Rahmen der Beurteilung, was angemessen ist und was nicht.

Beispiel

Schauen Sie sich die vom hessischen Bildungsserver bereitgestellten Instrumente zur Einschätzung der Kompetenzen im Bereich sozial-emotionale Entwicklung und zum Arbeitsverhalten an.

Sehen Sie sich die Rubrik „Interesse“ und die Rubrik „Kontaktverhalten“ an.

Betrachtet man die auf dem Bildungsserver zur Verfügung gestellten Beobachtungsbögen enthalten diese etliche normative Setzungen darüber, was von Schülern und Schülerinnen erwartet bzw. erwünscht ist. Teilweise werden im Sinne von Persönlichkeitstests Eigenschaften erhoben. Hier fragt sich, ob die Schule an dieser Stelle nicht deutlich über ihr Ziel hinausschießt:

Wer bestimmt beispielsweise, dass Schüler und Schülerinnen Interesse am Unterricht zeigen *müssen*, damit sie als engagiert gelten? Unterricht kann langweilig sein und die Schüler und Schülerinnen sind ja nicht freiwillig im Unterricht. Diszipliniertes (aber nicht vom Interesse bestimmtes) Lernen darf ja deshalb nicht als schlechter bewertet werden. Man kann argumentieren, dass hier lediglich Informationen gesammelt werden, die der Lehrkraft Rückschlüsse erlauben, ob zum Beispiel ihr Unterrichtsangebot interessant ist und zum Lernen motiviert. Die Formulierungen in den Beobachtungen sind aber deutlich auf eine Beurteilung und Bewertung der Beobachteten ausgelegt.

Weiter darf bezweifelt werden, dass die Einschätzung, ob ein Schüler „Unternehmensbereitschaft und Entschlussfähigkeit“ zeigt, für dessen

Einschätzung im Rahmen von Schule relevant ist. An anderen Stellen wiederum lassen die Formatierungen große Interpretationsspielräume zu: „Findet leicht Kontakt zu Lehrkräften“, „Findet leicht Kontakt zu Mitschülern/Mitschülerinnen“, „Sucht vorrangig gefühlsbezogene Zuwendung“. Wie ist „leicht“ definiert? Und was bedeutet *gefühlsbezogene Zuwendung*, und ist die Suche danach eine Schwäche? Warum ist es überhaupt ein Ziel „leicht Kontakt zu finden“? Und hängt dieses nicht auch maßgeblich von der Zusammensetzung der jeweiligen Gruppe ab?

Es eröffnet sich hier insgesamt das Problem einer normierten Schülerpersönlichkeit: Die allumfassenden Listen, die (fast) jeden Bereich der Persönlichkeitsentwicklung thematisieren, scheinen ein Ziel vorzugeben, in dessen Richtung sich die Schüler und Schülerinnen zu entwickeln haben. Wir haben es aber mit Persönlichkeiten zu tun, die selbstverständlich auch das Recht auf freie Entfaltung haben. Entsprechend werden sie natürlich auch unterschiedliche Interessen haben, mal mehr mal weniger motiviert sein, introvertierter oder extravertierter sein. Die Gefahr besteht, dass eine “ideale Schülerpersönlichkeit” konstruiert wird, die im Sinne einer Messlatte für die Beurteilung des Verhaltens genutzt wird. Dies soll im folgenden Beispiel verdeutlicht werden.

Beispiel

Der zwölfjährige Leonard fällt dadurch auf, dass er in Konfliktsituationen vermittelnd eingreift und sich auch bei Problemen seiner Mitschüler und Mitschülerinnen empathisch zeigt. Dies wird ihm positiv zurückgemeldet. Gleichzeitig jedoch erhält er die Rückmeldung, dass er zu empfindlich sei und auf schulische Situationen in denen er etwas nicht versteht, mit deutlicher Besorgnis reagiert. Auch kann es passieren, dass er, wenn ihm etwas nicht gelingt, Tränen in den Augen hat. Dieses wird ihm als unangemessene, übertriebene Reaktion zurückgemeldet.

Anhand des Beispiels wird deutlich, dass hier einmal ein deutlich erwünschtes Verhalten gelobt, ein vermeintlich unangepasstes/unangemessenes Verhalten kritisiert wird. Dabei wird übersehen, dass diese möglicherweise nicht unabhängig voneinander bestehen: An Leonards Fähigkeit sich empathisch zu zeigen und sensibel auf Konfliktsituationen zu reagieren, zeigt sich seine Sensibilität. Er nimmt Störungen, Stimmungen etc. deutlich wahr, versucht diese zu verstehen und reagiert darauf. Diese Sensitivität führt auch dazu, dass er bei selbstbezogenen Problemen empfindlicher reagiert als möglicherweise andere Personen und Gefühlsreaktionen zeigt, die andere möglicherweise nicht zeigen.

Zunächst kann man sich fragen, was überhaupt daran schlimm ist, empfindlicher zu sein als andere oder Gefühle zu zeigen. Diese schaden ja de facto niemandem. Darüber hinaus muss man sich klarmachen, dass eine Persönlichkeit viele Facetten und Ausprägungen hat, die sich möglicherweise bedingen. Oder anders ausgedrückt: Wäre es nicht besser das eine zu wertschätzen ohne das andere zu kritisieren? Warum *muss* eine Person auf sämtlichen Skalen den Normwert erreichen?

Es fehlen bislang empirische Untersuchungen, die prüfen, inwiefern sich explizite Einschätzungen zur Persönlichkeit von Schülern und Schülerinnen durch Lehrkräfte auf deren Leistungseinschätzung in anderen Bereichen auswirkt. Man darf zu Recht annehmen, dass das eine mit dem anderen zusammenhängt, da diverse Studien bereits den Einfluss von Schülermerkmalen auf die

Leistungsbewertung von Lehrkräften zeigen (Krämer & Zimmermann, 2020, Kaiser et al., 2013 Kaiser et al., 2015, Kaiser et al., 2017).

Verbale Beurteilung

Verbale Beurteilungen finden häufig bei der Rückmeldung zu mündlichen Leistungen im Rahmen des Unterrichts, nach Referaten oder auch begleitend im Rahmen der Rückgabe von Klassenarbeiten statt.

Verbale Beurteilungen, die vor der ganzen Klasse erteilt wurden, bergen verschiedene Probleme: Schüler und Schülerinnen vergleichen sich unmittelbar mit den anderen Schülern und Schülerinnen. Die öffentliche Beurteilung kann sich negativ auf das Selbstkonzept der Beurteilten auswirken und deren Status innerhalb der Klasse. Dies gilt sowohl für positive als auch für negative Rückmeldungen.

Eine Lösung ist die verbale Rückmeldung im geschützten Rahmen. Hier besteht die Möglichkeit auch kritische Punkte anzusprechen. Wiederum gilt, dass auch die verbale Rückmeldung wertschätzend, nachvollziehbar und am konkreten Verhalten orientiert sein soll. Die Lehrkraft muss sich auch hier wie bei allen anderen Einschätzungen und Rückmeldungen fragen, ob sie tatsächlich hinreichend viele diagnostische Situationen genutzt hat, um das Verhalten und die gezeigte Leistung zuverlässig beschreiben zu können.

Ein Problem der verbalen Rückmeldung liegt darin, dass gesprochene Worte naturgegeben flüchtig sind, es Missverständnisse und Erinnerungsfehler geben kann. Dies kann auch dazu führen, dass Eltern beispielsweise einen anderen Eindruck von der Einschätzung ihrer Kinder bekommen, wenn sie die Rückmeldung vermittelt über den Bericht ihrer Kinder erhalten. Die Kombination von schriftlicher und mündlicher Rückmeldung erscheint daher sinnvoll.

Zusammenfassend lässt sich feststellen, dass eine gute und sinnvolle Rückmeldung zunächst von der Qualität der vorangegangenen Diagnostik abhängt.

Weiter muss sie informativ, wertschätzend und auf das Verhalten bezogen sein, damit Schüler und Schülerinnen aus dieser etwas lernen können. Die auf der Grundlage der Rückmeldung erstellten Maßnahmen sollten sinnvollerweise alle Beteiligten im Blick behalten. Die einseitige Zuschreibung der Verantwortung auf nur einen oder wenige Akteure verhindert langfristig, dass Entwicklung sinnvoll gefördert werden kann. Und nicht zuletzt sollte man selbst immer wieder kritisch

reflektieren, ob die Beurteilung den Gütekriterien in möglichst hohem Maße genügen kann und ob die Rückmeldung angemessen und verständlich war.

LITERATUR

Arnold, K.-H. (2001). Qualitätskriterien für die Messung von Schulleistungen. In: Weinert, F.E. (Hrsg.): Leistungsmessungen in Schulen (S. 117-130). Weinheim: Beltz.

Baumert, J.; Artelt, C.; Klieme, E. & Stanat, P. (2001). PISA. Programme for International Student Assessment. Zielsetzung, theoretische Konzeption und Entwicklung von Messverfahren. In: Weinert, F.E. (Hrsg.): Leistungsmessungen in Schulen (S. 285-310). Weinheim: Beltz.

Birkel, C. & Birkel, P. (2002). Wie einig sind sich Lehrer bei der Aufsatzbeurteilung? Eine Replikationsstudie zur Untersuchung von Rudolf Weiss. In Psychologie in Erziehung und Unterricht, 49 (3), S. 219-224.

Birkel, P. (1978). Mündliche Prüfungen. Zur Objektivität und Validität der Leistungsbeurteilung. Bochumg: Kamp.

Birkel, P. (2005) Beurteilungsübereinstimmung bei Mathematikarbeiten? JMD 26, 28–51. https://doi.org/10.1007/BF03339005

Bourdieu, P. (1992). Ökonomisches Kapital - Kulturelles Kapital - Soziales Kapital. In: Pierre Bourdieu (Hrsg.): Die verborgenen Mechanismen der Macht. Schriften zu Politik & Kultur, 1. VSA, Hamburg, S. 49–79.

Carter, R.S. (1995). Wie gültig sind die durch Lehrer erteilten Zensuren? In: Ingenkamp, K. (Hrsg.): Die Fragwürdigkeit der Zensurengebung: Texte und Untersuchungsberichte (S.148-158). Weinheim: Beltz.

Deutsches Institut für Medizinische Dokumentation und Information (DIMDI) (Hrsg.) (2005). *Internationale Klassifikation der Funktionsfähigkeit, Behinderung und Gesundheit (ICF).* WHO, Genf

Dweck, C. S. & Leggett, E. L. (1988). A social cognitive approach to motivation and personality. *Psychological Review*, *95*, 256–273.

Elliot, A. J. & McGregor, H. A. (2001). A 2 x 2 achievement goal framework. *Journal of Personality and* Social Psychology, 80, 501–519.

Fiske, S. T. & Neuberg, S. L. (1990). A continuum of impression formation, from category based to individuating processes: Influences of information and motivation on attention and interpretation. Advances in Experimental Social Psychology, 23,1–74. https://doi.org/10.1016/s0065-2601(08)60317–2

Furck, C.L. (1975). Das pädagogische Problem der Leistung in der Schule (5. Aufl.). Weinheim: Beltz.

Gage, N.L. & Berliner, D.C. (1996). Pädagogische Psychologie. Weinheim: Psychologie Verlags Union.

Heyder, A.; Kessels. U. & Retelsdorf, J. (2019). Geschlechterstereotype in der Schule. In Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie (2019), 51, pp. 69-70. https://doi.org/10.1026/0049-8637/a000209. © 2019 Hogrefe Verlag.

Hofstätter, P.R. (1957). Tests. In: Fischer-Lexikon: Psychologie (S. 309-317). Frankfurt: Fischer.

Ingenkamp, K. & Lissmann, U. (2008). Lehrbuch der Pädagogischen Diagnostik. 6. Aufl. Weinheim: Beltz.

Jäger, R.S. (2007). Beobachten, beurteilen und fördern! Lehrbuch für die Aus-, Fort- und Weiterbildung. Landau: Verlag empirische Pädagogik.

Kaiser, J., Möller, J., Helm, F. & Kunter, M. (2015). Das Schülerinventar: Welche Schülermerkmale die Leistungsurteile von Lehrkräften beeinflussen. Zeitschrift für Erziehungswissenschaft, 18, 279 – 302.

Kaiser, J., Retelsdorf, J., Südkamp, A. & Möller, J. (2013). Achievement and engagement: How student characteristics influence teacher judgments. Learning and Instruction, 28, 73 – 84.

Kaiser, J., Südkamp, A. & Möller, J. (2017). The Effects of Student Characteristics on Teachers' Judgment Accuracy: Disentangling Ethnicity, Minority Status, and Achievement. Journal of Educational Psychology, 109, 871 – 888.

Katzenbach, D. / Schroeder, J. (2007). Ohne Angst verschieden sein zu können -Über Inklusion und ihre Machbarkeit. In: Zeitschrift für Heilpädagogik, 58:6, S. 202-213 [Downloadmöglichkeit unter www.inklusion-online.net]

Kemper, M. (2011). Normalisierung und Normalisierungskritik in der Pädagogik. Diskussionsbericht des 46. Salzburger Symposions. Vierteljahrsschrift für wissenschaftliche Pädagogik. 87 (2011) 4 - p. 599-617; link: https://doi.org/10.1163/25890581-087-04-90000002

Klauer, K. J. (2001). Wie misst man Schulleistungen. In: Weinert, F.E. (Hrsg.): Leistungsmessungen in Schulen (S. 103-116). Weinheim: Beltz.

Krämer, S. & Zimmermann, F. (2020). Zum Einfluss von störendem Schülerverhalten im Unterricht auf Leistungsbeurteilungen: Explizite Einschätzungen und experimentelle Befunde. In Zeitschrift für Pädagogische Psychologie Mär 2020, Vol. 34, Issue 2, S. 99-115

Lautmann, R. (1971). Gesellschaftliche Mechanismen im Examen. In: Eckstein, B. (Hrsg.): Hochschulprüfungen: Rückmeldung oder Repression? Blickpunkt Hochschuldidaktik Nr. 13, S.35-41.

Lienert, G.A. & Raatz, U. (1998). Testaufbau und Testanalyse (6. Aufl.). Weinheim Psychologie Verlags Union.

Moeller, M.L. (1972). Zur Psychodynamik des Prüfungswesens. Zeitschrift für Psychotherapie und Psychologie, 22, 1 1-13.

Paradies, L.; Wester, F. & Greving, J. (2005). Leistungsmessung und –bewertung. Berlin: Cornelsen.

Pawlik, K. (Hrsg.) (1976): *Diagnose der Diagnostik*. Stuttgart: Klett, 1976. 2. Aufl. 1982. Span.: Barcelona: Herder.

Preiser, Siegfried (2003). Pädagogische Psychologie. Psychologische Grundlagen von Erziehung und Unterricht (S.57). Weinheim und München: Juventa Verlag (Stangl, 2020).

Rheinberg, F. (2001). Bezugsnormen und schulische Leistungsbeurteilung. In: Weinert, F.E. (Hrsg.): Leistungsmessungen in Schulen (S. 59-72). Weinheim: Beltz.

Rindermann, H. (2008). International vergleichende Schulleistungs- und Intelligenzstudien: Worauf sind die Unterschiede zwischen Staaten zurückführbar? Versuch einer Erklärung unter ausschließlicher Berücksichtigung von Bildungsmerkmalen. Empirische Pädagogik. 22 (2008) 1 - p. 17-48 , 2008

Rost, D.H. & Scherner, F.J. (2001). Leistungsängstlichkeit. In: Rost, D.H. (Hrsg.): Handwörterbuch Pädagogische Psychologie (S. 405-413). Weinheim: Beltz.

Ståhlberg, J.; *Tuominen*, H.; Pulkka, A.-T. & Niemivirta; M. (2019) Maintaining the self? Exploring the connections between students' perfectionistic profiles, self-worth contingency, and achievement goal orientations. Personality and Individual Differences, Vol. 151

Stemmler, G. & Margraf-Stiksrud, J. (Hrsg.) (2015). Lehrbuch Psychologische Diagnostik. Hogrefe: Göttingen.

Towfigh, E.; Traxler, Ch. & Glöckner, A. (2014). Zur Benotung in der Examensvorbereitung und im ersten Examen. Zeitschrift für Didaktik der Rechtswissenschaft, 1, S. 8-27

Tuominen-Soini, H., Salmela-Aro, K. & Niemivirta, M. (2008). Achievement goal orientations and subjective well-being: A person-centred analysis. Learning and Instruction,18, 251–266

Twellmann, W. (Hrsg.) 1981: Handbuch Schule und Unterricht. Band VI. Düsseldorf: Schwann.