UNIVERSITÄT
BERN

Virtueller Campus Erziehungswissenschaft

Abteilung Pädagogische Psychologie

Pädagogische Diagnostik

Leitung: Prof. Dr. Walter Herzog

Autorenteam: Walter Herzog, Armin Hollenstein, Tamara Carigiet, Daniel Ingrisani, Elena Makarova, Hansjörg Lauener, Madeleine Pfäffli, Hans Ryser & Peter Vetter

Unter Mitwirkung von: Sonja Beeli, Lea Beyeler, Robert Hilbe, Julia Ignaczewska, Katrin Müller, Roman Suter, Belinda Vogt

Universität Bern
Virtueller Campus Erziehungswissenschaft
Institut für Erziehungswissenschaft
Abteilung Pädagogische Psychologie
Muesmattstrasse 27
3012 Bern
www.edu.unibe.ch

6. Auflage 2012

Inhaltsverzeichnis

Einleitung in die Pädagogische Diagnostik 1

Gegenstand pädagogischer Diagnostik 1
Beurteilung als professionelle Aufgabe 2
Zur Geschichte der pädagogisch-psychologischen Diagnostik 5
Qualitätskriterien von Prüfungen 7
Beurteilung als sozialer Prozess 10

Überblick über das Skript 12

1 Methodische Grundlagen der pädagogischen Diagnostik 13

1.1 Messen und Messtheorie 13
1.1.1 Messung psychologischer Merkmale und Eigenschaften 15

1.2 Skalentypen 18
1.2.1 Nominalskala 18
1.2.2 Ordinal- oder Rangskala 19
1.2.3 Intervallskala 19
1.2.4 Verhältnis- oder Proportionalskala 20

1.3 Grundzüge der klassischen Testtheorie 21

1.4 Gütekriterien 23
1.4.1 Objektivität 23
1.4.2 Reliabilität 25
1.4.3 Validität 26

1.5 Normen 28

2 Schulische Leistungsdiagnostik und Notengebung 32

2.1 Fallbeispiel 32

2.2 Schulische Leistungsbeurteilung 32
2.2.1 Der schulische Leistungsbegriff 33
2.2.2 Leistung zwischen Erinnern und Denken 34
2.2.3 Leistungsbeurteilung – Verpflichtung, Gegenstand und Form 34
2.2.4 Leistungsfeststellung, -bewertung und -rückmeldung 35
2.2.5 Funktionen der schulischen Leistungsbeurteilung 36

2.3 Diagnosekompetenz von Lehrpersonen 38
2.3.1 Was ist Diagnosekompetenz? 39
2.3.2 Empirischer Forschungsstand 40

2.4 Nicht-standardisierte Verfahren zur schulischen Leistungsdiagnostik 42
2.4.1 Mündliche Prüfungen 43

2.4.2 Aufsatzprüfung 45
2.4.3 Selbstbeurteilung 45
2.4.4 Portfolio 47

2.5 Notengebung in der Schule 48
2.5.1 Kritik an der Notengebung 49
2.5.1.1 Subjektive Fehlerquellen 49
2.5.1.2 Das klassen- und fachinterne Bezugssystem 50
2.5.1.3 Fazit 51
2.5.2 Verbesserungsmöglichkeiten der schulischen Leistungsbeurteilung 52
2.5.2.1 Inferenz- und Urteilsfehler vermindern 52
2.5.2.2 Transparente Leistungsbeurteilung und Notengebung 53
2.5.2.3 Einsatz standardisierter überregionaler Schulleistungstests 53
2.5.2.4 Förderorientierte und selektionsorientierte Notengebung 55
2.5.2.5 Fazit und Ausblick 57
2.5.3 Rehabilitierung der Notengebung 58

2.6 Standardisierte Schulleistungstests 60
2.6.1 Standardisierte bezugsgruppenorientierte Schulleistungstests 60
2.6.2 Standardisierte kriteriumsorientierte Schulleistungstests 62
2.6.3 Klassencockpit – Darstellung eines ausgewählten standardisierten Verfahrens 62
2.6.3.1 Hintergrund 62
2.6.3.2 Anwendung 63
2.6.3.3 Auswertung 63
2.6.3.4 Interpretation und Nutzung der Ergebnisse 65
2.6.3.5 Gütekriterien zum Klassencockpit 66
2.6.3.6 Fazit 66

2.7 Basic-Check und Multi-Check: Leistungsbeurteilung durch ausserschulische Instanzen 67

2.8 Grenzen der schulischen Leistungsdiagnostik 69

3 Begabungsdiagnostik 70

3.1 Fallbeispiel 70
3.1.1 Hochbegabung und Schulerfolg 71
3.1.2 Intelligenz als Kernbereich der Hochbegabung 72

3.2 Diagnostik der Hochbegabung 73
3.2.1 Nicht-standardisierte Verfahren 73
3.2.2 Standardisierte Verfahren 74
3.2.2.1 Kreativitätstest 74
3.2.2.2 Selbstbeurteilungstests 74
3.2.2.3 Fremdbeurteilungstests 74
3.2.2.4 Psychometrische Verfahren 75
3.2.2.5 Einzelintelligenztest 75
3.2.2.6 Eindimensionale Intelligenztests 75
3.2.2.7 Mehrdimensionale Intelligenztests 75

3.3 Intelligenztest: HAWIK-III 75
3.3.1 Verbalteil 77

3.3.2 Handlungsteil 79
3.3.3 Interpretation der Testergebnisse 81
3.3.4 Gütekriterien des Intelligenztestes HAWIK-III 82

3.4 Anwendungsmöglichkeiten: Intelligenztests bei Lernstörungen ... 82
3.4.1 Legasthenie 83
3.4.2 Dyskalkulie 84

3.5 Kritik am Einsatz von Intelligenztests ... 84
3.5.1 Kritik am Einsatz von Intelligenztests im Zusammenhang mit Hochbegabung 84
3.5.2 Kritik an den Intelligenztests 85

3.6 Fazit ... 86

4 Unterrichtsbeobachtung ... 87

4.1 Fallbeispiel Mobbing in der Klasse ... 87

4.2 Verhaltensbeobachtung ... 88
4.2.1 Beobachtung 88
4.2.2 Unterschiede zwischen unsystematischer und systematischer Beobachtung 89
4.2.3 Methodische Aspekte des Beobachtens 90

4.3 Formen der Beobachtung ... 90
4.3.1 Teilnehmende vs. nicht-teilnehmende Beobachtung 91
4.3.2 Offenes vs. verdecktes Beobachten 91
4.3.3 Standardisierte Beobachtung vs. unstandardisierte Beobachtung 91
4.3.4 Kontinuierliche vs. diskontinuierliche Beobachtungen 92
4.3.5 Labor- vs. Feldbeobachtung 92
4.3.6 Medial vermitteltes vs. unvermitteltes Beobachten 92
4.3.7 Zusammenfassung 92

4.4 Formen der Beobachtungsprotokollierung ... 93
4.4.1 Protokolltechniken 93
4.4.1.1 Gedächtnisprotokoll 93
4.4.1.2 Strichlisten 94
4.4.1.3 Wortprotokoll 95
4.4.1.4 Teilformalisiertes Protokoll 96
4.4.2 Technische Hilfsmittel 96

4.5 Beobachtungssysteme ... 97
4.5.1 Überblick über die Beobachtungssysteme 97
4.5.1.1 Zeichensysteme 97
4.5.1.2 Kategoriensysteme 98
4.5.1.3 Ratingskalen 100
4.5.1.4 Soziogramme 101

4.6 Anwendungsmöglichkeiten der Beobachtung ... 104
4.6.1 Allgemeine Aspekte 104
4.6.2 Anwendungsmöglichkeiten des dargestellten Beispiels 106

4.7 Grenzen der Unterrichtsbeobachtung 107
4.7.1 Allgemeine Grenzen der Beobachtungsmethode 108
4.7.2 Grenzen der Unterrichtsbeobachtung am Beispiel Mobbing 109

4.8 Fazit 110

5 Verzeichnisse 111

5.1 Literaturverzeichnis 111

5.2 Internetseiten 118

Einleitung in die Pädagogische Diagnostik

Was verstehen wir unter Diagnostik? Vielleicht assoziieren wir einen Arztbesuch. Oder einen Anruf bei der Pannenhilfe. Der Arzt soll herausfinden, weshalb wir uns in letzter Zeit so müde fühlen. Die Pannenhilfe soll feststellen, warum der Motor unseres Fahrzeugs nicht anspringen will. Vielleicht denken wir auch an einen Intelligenz- oder Interessenstest, den wir bei einem Berufsberater gemacht haben, der uns bei der Berufswahl beraten hat. Als Naturfreunde kommt uns vielleicht die Bestimmung einer Pflanzen- oder Tierart in den Sinn. Oder wir verbinden Diagnose mit Prognose und erinnern uns mit Schrecken an den Arzt, der uns aufgrund seines Urteils über die vermeintlichen Ursachen unserer Müdigkeit wenig Hoffnung auf ein langes Leben machte.

Diagnose hat mit Beurteilung zu tun, mit der Analyse eines Sachverhalts, den wir auf Anhieb nicht zu verstehen oder zu erkennen vermögen, der deshalb einer Sondierung bedarf, um aufgeklärt zu werden. Gemäss Duden heisst Diagnose «unterscheidende Beurteilung» (Bibliographisches Institut & F.A. Brockhaus 2001, p. 220). Zu solcher «unterscheidender Beurteilung» sind wir in verschiedener Hinsicht herausgefordert: in Bezug auf die Natur (Bestimmung von Pflanzen und Tieren), in Bezug auf die Technik (Fehlerdiagnose bei Geräten oder Fahrzeugen), in Bezug auf unseren Körper (Diagnose des Arztes) und in Bezug auf unsere Psyche (Interessens- und Intelligenztest des Berufsberaters). Überall, wo nicht offen zutage liegt, womit wir es zu tun haben, sind wir auf diagnostische Urteile angewiesen. Das gilt auch und nicht zuletzt im Kontext von Erziehung und Unterricht. Die *pädagogische Diagnostik* befasst sich mit der «unterscheidenden Beurteilung», die Personen vollziehen, die in pädagogischen Kontexten tätig sind.

In diesem einleitenden Kapitel wollen wir einen kursorischen Einblick in Gegenstand, Aufgaben, Verfahren und Kriterien der pädagogischen Diagnostik geben. Dabei werden wir in einem ersten Durchgang Fragen wie die folgenden ansprechen: Worum geht es bei der pädagogischen Beurteilung? Inwiefern stellt die Beurteilung eine professionelle Aufgabe von Lehrerinnen und Lehrern dar? Was sind die historischen Wurzeln der pädagogischen Diagnostik? Welche Qualitätskriterien sind bei diagnostischen Urteilen zu beachten? Weshalb ist die Beurteilung anderer Menschen ein sozialer Prozess?

Gegenstand pädagogischer Diagnostik

Das breite Feld diagnostischen Verhaltens, das wir einleitend abgesteckt haben, wird im Kontext von Erziehung und Unterricht im Wesentlichen auf psychologische und soziale Sachverhalte begrenzt. Es geht also um die Beurteilung von Menschen und von Beziehungen zwischen Menschen. Die Beurteilung anderer Menschen ist ein Vorgang, den wir täglich, ja oft stündlich vollziehen. In sozialen Situationen sind wir darauf angewiesen zu erfahren, was die andere Person meint, was sie beabsichtigt, wer sie ist, ob wir ihr vertrauen können, weshalb sie uns anspricht, welche Meinungen sie vertritt, was sie von uns hält, ob sie uns freundlich oder feindlich gesinnt ist, warum sie uns sympathisch ist etc. Oft geht es bei der Beurteilung anderer Menschen um ganz banale Dinge wie die Feststellung des Geschlechts: Haben wir es mit einem Mann oder mit einer Frau zu tun? Oder des Berufs: Welche berufliche Tätigkeit übt die Person aus? Oder des Alters: Wie alt ist sie? Oder der Herkunft: Ist sie von hier oder von auswärts? Solche äusseren Merkmale wie Geschlecht, Alter, Beruf und Herkunft sind oft ausreichend, um unser Gegenüber zu beurteilen und unser Verhalten ihm gegenüber zu gestalten.

Die angemessene Beurteilung anderer Menschen ist eine Voraussetzung für gelingende soziale Interaktionen und Beziehungen. Ohne eine minimale Kenntnis der anderen Person ist es uns nicht möglich, ihr gegenüber richtig zu handeln. So fällt es uns ausserordentlich schwer, mit einer ande-

ren Person zu interagieren, wenn wir nicht fähig sind, deren Geschlecht oder Alter zu eruieren. Die Beurteilung ist aber nicht nur eine *Voraussetzung* des sozialen Umgangs, sie ist auch eine Tätigkeit, die wir im Allgemeinen fast automatisch und *ohne grosse Probleme* vollziehen. Menschen verfügen über eine intuitive Kompetenz zur Beurteilung anderer, die es ihnen erlaubt, sich situativ richtig zu verhalten: Mit einer Frau gehen wir anders um als mit einem Mann, den Briefboten behandeln wir neutraler als den Anlageberater, gegenüber einem Fremden benehmen wir uns vorsichtiger als gegenüber dem Nachbarn etc. Unsere Urteile über andere Menschen sind *funktional* für unsere sozialen Beziehungen und funktionieren im Allgemeinen ganz gut.

Beurteilung als professionelle Aufgabe

Das gilt auch für pädagogische Situationen. Allerdings sind die Ansprüche an unser Urteilsvermögen in pädagogischen Kontexten im Allgemeinen höher als im alltäglichen Umgang – jedenfalls dann, wenn wir Lehrerin oder Lehrer sind.[1] Von Menschen, die einen Beruf ausüben, der wesentlich darauf beruht, dass sie über andere Menschen urteilen, wird zu Recht erwartet, dass sie ihre Urteile nicht nur intuitiv fällen, sondern über eine professionelle Ausbildung zur Beurteilung anderer verfügen. Dazu kommt, dass im Kontext der Schule «äusserliche» Merkmale wie Geschlecht, Alter und Herkunft zwar auch eine Rolle spielen, aber wichtiger sind weniger leicht erkennbare, «innere» Merkmale wie die Intelligenz von Schülerinnen und Schülern, ihre Begabungen und Interessen, ihre Kenntnisse, ihre Lernstrategien, ihr leistungsbezogenes Selbstbild, ihre Persönlichkeit etc. Denken wir auch an Schülerinnen und Schüler mit körperlichen oder psychischen Defiziten (u.a. ADS bzw. ADHS[2], Sehbehinderung, motorischen Störungen, Legasthenie, Underachiever), aber auch an solche mit besonderen Begabungen (u.a. Hochbegabung, musikalische oder sportliche Talente). Denken wir auch an die ganz gewöhnlichen Fälle eines durchschnittlichen Drittklässlers oder einer durchschnittlichen Schülerin der zweiten Sekundarstufe. Die Lehrperson muss fähig sein, auch diesen durchschnittlichen Schüler und diese durchschnittliche Schülerin in verschiedener Hinsicht adäquat zu beurteilen: im Hinblick auf allfällige Wissenslücken, Stärken und Schwächen, Entwicklungspotenzial, Förderungsmöglichkeiten etc. Auch über die soziale Position des Schülers bzw. der Schülerin in der Klasse sollte die Lehrperson im Bilde sein, über den familiären Hintergrund, weshalb er oder sie in letzter Zeit so niedergedrückt wirkt oder warum er oder sie bei der letzten Probe so schlecht abgeschnitten hat.

Um die Beurteilungsaufgaben einer Lehrperson etwas systematischer herzuleiten, kann man sich an den Aufgaben orientieren, die Lehrerinnen und Lehrer im Berufsfeld Schule zu erfüllen haben. Da ist als Erstes die *pädagogische* Aufgabe der Schule: Es wird erwartet, dass die Schule den einzelnen Schüler und die einzelne Schülerin fördert, damit sie zu sich selber finden und ihre Begabungen und Interessen sowie ihre Persönlichkeit entfalten können. Dies setzt voraus, dass die Lehrerin bzw. der Lehrer fähig ist, Begabungen und Interessen sowie Persönlichkeiten adäquat zu erkennen: Was ist das Potenzial *dieser Schülerin oder dieses Schülers*? Wie intelligent ist sie oder er? Wofür ist sie oder er begabt? Welche Persönlichkeit hat sie oder er? Wie kann sie oder er gefördert werden? Was können Lehrpersonen tun, um ihre Schülerinnen und Schüler in ihrer Entwicklung zu unterstützen? Aus der pädagogischen Aufgabe geht die Notwendigkeit einer *Förderdiagnostik* hervor: Lehrperso-

---

[1] Von den Schülerinnen und Schülern wird nicht erwartet, dass sie über eine hohe Beurteilungskompetenz verfügen, obwohl sie oftmals erstaunlich gut in der Lage sind, ihre Lehrpersonen zu beurteilen. Des Weiteren kann es ein Ziel des Unterrichts sein, die Kompetenz der Schülerinnen und Schüler zur *Selbst*beurteilung zu stärken.

[2] ADS oder ADHS ist die deutsche Bezeichnung für das englische «Attention-Deficit-Disorder» (ADD) und bezeichnet eine international anerkannte psychiatrische Diagnose. «Diese ist gekennzeichnet durch seit der Kindheit bestehende und ausgeprägte Störungen der Konzentrationsfähigkeit, der Planungs- und Handlungskontrolle, durch Störungen der Impulskontrolle sowie durch motorische Hyperaktivität» (www.adhs.ch).

nen müssen fähig sein, ihre Schülerinnen und Schüler so zu beurteilen, dass sie ihnen die bestmögliche Förderung zukommen lassen können.

Eine zweite Aufgabe der Schule liegt in der *Wissensvermittlung*: Schule soll Wissen, Kenntnisse und Fertigkeiten vermitteln, dies im Sinne der Tradierung des kulturellen Bestandes einer Gesellschaft. Diese Aufgabe der Stoffvermittlung setzt voraus, dass sich die Lehrperson über verschiedene «innere» Bedingungen ihrer Klientel ins Bild zu setzen vermag: Was wissen die Schülerinnen und Schüler bereits? Wo kann ich sie «abholen»? Was ist ihr Vorwissen über den Stoff, den ich in dieser Lektion durchnehme? Über welche Fehlauffassungen und Vorurteile verfügen sie? Sind alle auf dem gleichen Wissensstand? Welches ist die beste Methode, um ihnen den neuen Stoff beizubringen? Zur Aufgabe der Wissensvermittlung gehört auch die Prüfung des *Wissenszuwachses*: Haben die Schülerinnen und Schüler etwas gelernt? Was haben sie gelernt? Wie viel haben sie gelernt? Ist ihnen das Lernen leicht oder schwer gefallen? Wie sehr haben sie sich angestrengt? Wer hat mehr, wer hat weniger gelernt? Habe ich das vom Lehrplan vorgegebene Ziel erreicht?

Eine dritte Aufgabe der Schule liegt in der Klassifikation bzw. *Selektion* der Schülerinnen und Schüler nach ihrem Leistungsgrad. Schule, so wie wir sie kennen, ist ein soziales Sieb: Sie soll darüber befinden, wer nach Abschluss einer Bildungsetappe die nächste Etappe in Angriff nehmen darf und wer durch «Sitzenbleiben» die Etappe wiederholen muss, und sie soll festlegen, wer nach Abschluss einer längeren Bildungsphase in welchen Typus von weiterführender Schule wechseln darf. Für die Lehrperson stellt sich die Frage, ob ihre selektionsrelevanten Entscheidungen korrekt sind: Sind die Prüfungen, die ich durchführe, verlässlich und gerecht? Stelle ich die richtigen Fragen? Bin ich unvoreingenommen bei meiner Korrekturarbeit? Wie soll ich überhaupt prüfen: Indem ich offene oder geschlossene (Multiple-Choice-)Fragen stelle? Indem ich vorweg festlege, welche Leistungen welche Noten ergeben, oder indem ich zuerst schaue, wie sich die Leistungen in der Klasse verteilen und erst dann bestimme, wo die Grenze zwischen «genügend» und «ungenügend» liegt? Um der Aufgabe der Selektion gewachsen zu sein, bedarf eine Lehrperson der Kompetenz zur *Leistungsdiagnostik*.

Eine vierte Aufgabe der Schule liegt in der *Sozialisation*: Die Schule soll nicht nur Wissen vermitteln, nicht nur Selektion betreiben und nicht nur jedem zu sich selber verhelfen, sondern auch die zentralen Werte und Normen unserer Gesellschaft weitergeben. Wie beurteile ich den Erfolg meines Unterrichts in dieser Hinsicht? Wie erfasse ich das Sozialverhalten der Schülerinnen und Schüler? Wo liegt die Grenze zwischen aggressivem und nicht-aggressivem Verhalten, zwischen gutem und schlechtem Benehmen, zwischen echtem und geheucheltem Interesse, zwischen Aufschneiderei und legitimer Selbstdarstellung etc.?

Schliesslich sollten wir fünftens – auch wenn hier nicht von einer «Aufgabe» gesprochen werden kann – nicht vergessen, dass pädagogische, vor allem schulische Situationen *soziale* Situationen sind. Wir haben es im Allgemeinen nicht mit einem einzigen Schüler oder einer einzigen Schülerin zu tun, sondern mit einer ganzen Schulklasse. Als Lehrerin oder Lehrer sind wir darauf angewiesen, die soziale Dynamik dieser Schulklasse wenigstens so weit zu erkennen, dass wir störendes Verhalten unterbinden können. Aber vielleicht sollten wir auch etwas darüber wissen, wie wir die Gruppensituation konstruktiv für die Aufgaben der Stoffvermittlung und der Sozialisation nutzen können. Dazu genügt die Kompetenz zur Beurteilung von Individuen nicht. Wir müssen fähig sein, soziale Strukturen und soziale Abläufe zu erkennen – eine unter Umständen anspruchsvollere Aufgabe als die Diagnose der Kompetenzen, Motive und Eigenschaften von Individuen.

Dieser kurze Blick auf die Aufgaben der Schule und ihren Charakter als soziale Situation zeigt, wo die Ansprüche an eine *professionelle* Beurteilungskompetenz der Lehrpersonen liegen. Zwar sind wir als Menschen mit einem «natürlichen» Potenzial an Beurteilungskompetenz ausgestattet. Wir erkennen oft intuitiv, welche Absichten eine andere Person hat, welchen Charakter sie hat, wie viel Vertrauen wir ihr schenken können, weshalb wir den Umgang mit ihr meiden sollten etc. Diese In-

tuition mag völlig ausreichend sein, solange unsere Urteile keine nachhaltigen Konsequenzen für andere Menschen haben. Genau dies aber ist in pädagogischen Situationen der Fall: Hier geht es nicht um uns selbst, sondern um die Beurteilung anderer, und zwar mit Konsequenzen für deren Lebenschancen, für deren Zukunft und für deren persönliche Entwicklung. Unter diesen Umständen kann die Intuition nicht genügen; sie muss gestützt werden durch Hilfsmittel und professionelle Urteilskompetenz.

Was ermöglicht uns überhaupt Urteile über andere Menschen? Als Informationsquellen stehen uns die Beobachtung ihres Verhaltens, das Verstehen ihrer sprachlichen Äusserungen (Gespräche), die Deutung ihres nonverbalen Ausdrucks (Mimik, Gestik, Physiognomie u.ä.) und die Analyse von Objektivationen, das heisst Erzeugnissen ihrer Leistungskraft und Kreativität (Zeichnungen, Schriftstücke, Produkte irgendwelcher Art u.ä.), zur Verfügung. Bekanntlich kann man nicht in die Psyche oder das Bewusstsein eines anderen Menschen Einblick nehmen. Wir sehen nicht, was «im Kopf» des anderen vor sich geht. Niemand vermag sich im wörtlichen Sinn in eine andere Person zu versetzen. Immer beruhen unsere Urteile über andere Menschen auf Schlüssen, die wir aufgrund von Hinweisen, die wir ihrem Äusseren, ihrem Verhalten oder ihren Objektivationen entnehmen, ziehen. Psychisches erkennen wir aufgrund von Beobachtungen, die wir anstellen und anschliessend *interpretieren*.

Bei solchen Interpretationen sind Missverständnisse fast unausweichlich. Oft urteilen wir zu schnell: Jemand ist uns auf Anhieb sympathisch oder unsympathisch, wir empfinden ein Gegenüber als schön oder hässlich, wir fühlen uns zu einer Person hingezogen oder von ihr abgestossen. Solche ersten Eindrücke können sich mit «Halo-Effekten» verbinden: Wir assoziieren schön mit intelligent, intelligent mit gewissenhaft, gewissenhaft mit umgänglich, umgänglich mit kreativ etc. So ergibt sich aus dem ersten Eindruck ein fast komplettes Bild der Persönlichkeit der anderen Person. Doch das rasch angefertigte Bild kann falsch sein und das sympathische Gegenüber stellt sich bei genauerer Betrachtung als Egoistin oder Egoist heraus.

Was für den Alltag vielleicht weniger gravierend ist, weil unsere Fehlurteile Erwachsene betreffen, die sich gegen falsche oder unliebsame Schlüsse wehren können[3], ist in pädagogischen Kontexten unter allen Umständen zu vermeiden. Nichts kann schlimmer sein für einen Schüler oder eine Schülerin, als wenn er bzw. sie vorschnell in ein Schema gepresst wird, wenn ihm bzw. ihr ein Etikett angehängt wird oder wenn er bzw. sie aufgrund äusserer Merkmale oder einer kurzschlüssigen Deutung dieser Merkmale fehlbeurteilt wird. Zur Professionalität von Lehrpersonen gehört ganz wesentlich die Fähigkeit, diagnostische Urteile sorgfältig und reflektiert, das heisst im Bewusstsein der möglichen Fehler, die dabei unterlaufen können, vornehmen zu können.

Die wissenschaftliche Disziplin, die den Lehrpersonen helfen soll, bei ihren Beurteilungen Professionalität zu erreichen, wird *Pädagogische Diagnostik* genannt. Da die pädagogische Diagnostik sehr stark von den in der Psychologie entwickelten diagnostischen Verfahren beeinflusst ist, wird oft auch von *Pädagogisch-psychologischer Diagnostik* gesprochen. Wir unterscheiden die beiden Begriffe nicht.[4] Aufgabe der pädagogischen bzw. pädagogisch-psychologischen Diagnostik ist die Klärung der Probleme der diagnostischen Urteilsbildung in pädagogischen Feldern (insbesondere

---

[3] Allerdings dürfen wir uns keine Illusionen machen. Urteile über andere Personen kommen oft unreflektiert zustande und lassen sich von Seiten der Betroffenen nicht leicht korrigieren. In gravierenden Fällen kann aus Fehlurteilen Mobbing hervorgehen.

[4] Der Begriff der *pädagogischen* Diagnostik ist jüngeren Datums. Er stammt aus den 1970er Jahren, wenn auch frühere Verwendungsweisen durchaus nachgewiesen werden können. Die 1970er Jahre waren eine Zeit, als die Pädagogik sich den Methoden der empirischen Forschung öffnete und psychologische Verfahren für ihre eigenen Problemstellungen übernahm bzw. adaptierte. Im Englischen entspricht dem Begriff der Pädagogischen Diagnostik die Bezeichnung «Educational Measurement» (vgl. Süllwold 1983).

der Schule) sowie die Bereitstellung von Kriterien und Hilfsmitteln zur Optimierung des Prozesses der pädagogischen Urteilsbildung.

Zur Geschichte der pädagogisch-psychologischen Diagnostik

Historisch gesehen gibt es diagnostische Hilfsmittel spätestens seit den Typenlehren der Antike. Bekannt ist bis auf den heutigen Tag die Lehre von Galen (130-201 n. Chr.), einem Arzt, der vier Temperamente unterschieden hat: den Sanguiniker (sorglos und optimistisch), den Phlegmatiker (träge und verträglich), den Melancholiker (besorgt und schwermütig) und den Choleriker (aufbrausend und habsüchtig). Typologien haben eine gewisse Faszination, weil sie die Fülle an individuellen Differenzen zwischen den Menschen auf wenige Aspekte reduzieren und damit einen leichten und schnellen Überblick über andere gewinnen lassen. Sie sind aber mit Skepsis zu beurteilen, da Typen oft eher Karikaturen gleichen und selten mit konkreten Einzelpersonen in Übereinstimmung zu bringen sind. Auch beruhen die Typenlehren zumeist nicht auf wissenschaftlicher Forschung, sondern sind eher intuitiv begründet. Psychologische Typenlehren finden sich aber bis weit ins 20. Jahrhundert hinein – Beispiele sind die Lehre von den drei Körperbautypen (Pykniker, Athletiker und Leptosome) von Ernst Kretschmer (1888-1964), die Differenzierung in Extravertierte (nach aussen Gerichtete) und Introvertierte (nach innen Gerichtete) des Schweizer Psychiaters Carl Gustav Jung (1875-1961) und die Unterscheidung verschiedener menschlicher Lebensformen – theoretische, ökonomische, ästhetische, soziale, machtbezogene und religiöse Lebensform – des Pädagogen Eduard Spranger (1882-1963).

In der Psychologie sind seit dem Ende des 19. Jahrhunderts Verfahren entwickelt worden, die eine genauere und verlässlichere Analyse individueller Differenzen vornehmen lassen. Auf der einen Seite waren es Schüler von Wilhelm Wundt (1832-1920), dem Begründer der wissenschaftlichen Psychologie, die erste Ansätze *standardisierter* Verfahren zur Erfassung von individuellen Differenzen entwickelten – zu erwähnen sind insbesondere Emil Kraepelin (1856-1926), Hermann Ebbinghaus (1850-1909) und Ernst Meumann (1862-1915). Zum anderen war es der Engländer Francis Galton (1822-1911), der erste *Apparate* zur Messung individueller Differenzen entwickelte. Ihm zur Seite standen der Amerikaner James McKeen Cattell (1860-1944), auch er ein Schüler von Wundt, und der Franzose Alfred Binet (1857-1911). Letzterer hat den ersten Intelligenztest entwickelt, indem er Kindern verschiedenen Alters Aufgaben unterschiedlichen Schwierigkeitsgrades vorlegte und feststellte, inwiefern sie sich im Vergleich mit einer Gruppe altersgleicher Kinder in ihren Leistungen unterschieden. Daraus ist der Begriff des *Intelligenzquotienten* hervorgegangen, den allerdings nicht Binet, sondern der deutsche Psychologe William Stern (1871-1938) prägte. Definiert ist der Intelligenzquotient (IQ) als Verhältnis von Intelligenzalter zu Lebensalter multipliziert mit 100. Konkret: ein Kind von 7 Jahren, das – gemessen mit einer standardisierten Reihe von Intelligenzaufgaben – Leistungen erbringt, zu denen im Durchschnitt erst neunjährige Kinder fähig sind, hat einen IQ von (9 : 7) x 100 = 128.5.[5] Binet entwickelte seine Testreihen ursprünglich, um minderbegabte Kinder, die einer Sonderschule zugewiesen werden sollten, erkennen zu können.

Neben der Intelligenz wurden zu Beginn des 20. Jahrhunderts auch andere Merkmale von Kindern und Jugendlichen diagnostisch erfasst. So wurden Masse für die Konzentrationsfähigkeit ausgearbeitet sowie verschiedene Formen von Begabungs- und Leistungstests entwickelt (vgl. Schmid 1977). Ein Sonderproblem stellte die so genannte Überbürdungsfrage dar. Die zunehmend länger werdende Dauer der Beschulung der Kinder, wie sie gegen Ende des 19. Jahrhunderts einsetzte, liess die Frage aufkommen, ob dies den Kindern überhaupt gemäss sei oder ob sie nicht durch den

[5] Bei den heute üblichen IQ-Tests ist das Verfahren allerdings etwas anders. Bestimmt wird die Intelligenz nicht mehr als Quotient, sondern als Abweichung vom Mittelwert einer normierten Verteilung von Intelligenzleistungen, der bei 100 liegt. Man spricht deshalb von einem «Abweichungs-IQ» (vgl. Abschnitt 3.1.1). Im Alltag ist allerdings trotzdem weiterhin einfach vom IQ die Rede.

Schulbesuch überlastet (eben: überbürdet) würden. Man versuchte diese Frage durch die Entwicklung von Tests zu klären, deren Aufgabe die Erfassung von Ermüdung und Leistungsabfall war (vgl. Groffmann 1964; Schmid 1977, p. 97ff.).

Zu einem wichtigen Thema der pädagogisch-psychologischen Diagnostik wurde die *Schulreife*. Vor allem in den Jahren nach dem zweiten Weltkrieg wurden verschiedene Tests entwickelt, die darüber befinden sollten, ob ein Kind «reif» ist, um eingeschult zu werden. Später sind *Schulleistungstests* dazugekommen, die bezogen auf eine Altersgruppe oder eine Schulstufe normierte Aussagen über das Leistungsvermögen bzw. das Schulwissen eines Kindes oder Jugendlichen ermöglichen. In Deutschland gibt es eine ganze Reihe solcher Schulleistungstests, und zwar für fast alle Schulstufen (vgl. Ingenkamp 1981; Jäger 1992). Sie betreffen Bereiche wie Deutsch, Rechnen und Mathematik, Fremdsprachen, naturwissenschaftliche Fächer, aber auch Lesefertigkeiten, Rechtschreibefähigkeiten, Konzentrationsfähigkeit, Lernbehinderungen, Sozialverhalten, Hochbegabung u.a. Was die Schweiz anbelangt, so bestehen nur wenige Schulleistungstests, was im Wesentlichen mit der föderalistischen Struktur unseres Bildungswesens (kantonale Hoheit im Bildungsbereich, insbesondere bei den Lehrplänen) zu tun hat. Die Situation hat sich insofern verändert, als im Zusammenhang mit internationalen Schulleistungsvergleichsstudien wie TIMSS[6] und PISA[7] standardisierte Verfahren zur vergleichenden Beurteilung von Schülerleistungen auch bei uns Einzug halten.

Gegenüber standardisierten Wissenstests bestehen grundsätzliche Vorbehalte, weshalb eine eigentliche Pendelbewegung zu beobachten ist: einmal haben sie Hochkonjunktur, dann wieder werden sie Opfer einer Rezession. Zur Zeit erleben wir im Rahmen der eben erwähnten internationalen Vergleichsstudien, die in der Schweiz in das Projekt eines *nationalen Bildungsmonitoring* eingebettet sind, einen positiven Pendelausschlag.[8] Wie lange er anhalten wird, ist schwer zu sagen.

Es wäre falsch, die pädagogische Diagnostik mit *Tests* und anderen standardisierten Verfahren gleichzusetzen. Tests sind *eine* Methode, die uns helfen kann, den Beurteilungsprozess zu professionalisieren, aber keineswegs die einzige. Die messtheoretischen Schwächen der Tests haben auch schon früh dazu geführt, dass alternative Beurteilungsformen gefordert wurden. Alternativen können sein: Beobachtung des Schülerinnen- und Schülerverhaltens durch die Lehrperson (dabei werden oft andere Aspekte als nur das Lern- und Leistungsverhalten sichtbar), Beurteilungsgespräche, Portfolio, Selbstbeurteilung, Unterricht nach dem Prinzip des *Mastery Learning*[9], Wortgutachten (statt mittels Noten werden die Lernenden verbal beurteilt), Assessment-Verfahren etc. Letztlich gilt aber auch für solche alternativen Beurteilungsformen, dass sie in Hinblick auf ihre Objektivität, Reliabilität und Validität überprüft werden müssen, so dass sich die Vor- und Nachteile verschiedener Beurteilungssysteme *vergleichend* darstellen lassen.

Gegen Tests spricht auch, dass Lehrerinnen und Lehrer selten in der Lage sind, diese selber einzusetzen; im Vordergrund ihrer Beurteilungspraxis stehen andere Verfahren, wie insbesondere die Beobachtung des Schülerinnen- und Schülerverhaltens, das Gespräch mit Lernenden und die Beurteilung ihrer Leistungen. Letzteres betrifft im Kontext der Schule jede Art von manifester Leistung,

---

[6] TIMSS steht für «Trends in International Mathematics and Science Study» und ist eine international vergleichende Schulleistungsuntersuchung, die Mathematik- und Naturwissenschaftsleistungen in der Grundschule, in der Sekundarstufe I und II untersucht (vgl. www.timss.mpg.de).

[7] PISA steht für «Programme for International Student Assessment» und ist eine internationale, durch die OECD geschaffene Studie, welche darauf abzielt, die Kompetenzen von 15-jährigen Schülerinnen und Schülern zu messen (vgl. www.pisa.admin.ch).

[8] Vgl. www.edk.ch > Arbeiten > Bildungsmonitoring.

[9] *Mastery Learning* ist ein Begriff des amerikanischen Bildungsforschers Benjamin Bloom und meint, dass so unterrichtet wird, dass alle Schülerinnen und Schüler die gesetzten Lernziele erreichen (vgl. Bloom 1968). Im Deutschen wird Mastery Learning oft mit «zielerreichendem Lernen» übersetzt.

die ein Schüler oder eine Schülerin erbringt, wobei wir uns im Folgenden zum Zwecke der Illustration auf *Prüfungsleistungen* beziehen.

Qualitätskriterien von Prüfungen

Prüfungen haben die Funktion, den Leistungsstand und das Leistungsniveau eines Schülers oder einer Schülerin in einem Fach festzustellen. Wie jede Art von Beurteilung haben Prüfungen, wenn sie gewisse Qualitätsstandards erfüllen wollen, objektiv, reliabel und valide zu sein. Diese Kriterien gelten letztlich für jedes diagnostische Verfahren, also auch für Tests, Verhaltensbeobachtungen, Lernberichte, Portfolios etc. Wir werden diese Kriterien später noch eingehender erörtern. Im Folgenden geht es um ein erstes Vertrautwerden mit den Qualitätsansprüchen, die an die pädagogisch-psychologische Urteilsbildung aus einer professionellen Perspektive zu stellen sind, wobei wir als Beispiel die Notengebung herausgreifen.

Die gängige Notenskala (1 bis 6 mit halben Zwischenstufen wie 3.5 oder 5.5) erweckt den Eindruck eines exakten Messinstruments – vergleichbar der Skala eines Thermometers oder des Metermasses. Doch wie wir aus eigener Erfahrung wissen, kann die Note 5 in einem Fach etwas anderes bedeuten als dieselbe Note in einem anderen Fach. Selbst im gleichen Fach kann ein und dieselbe Note bei einer Lehrperson etwas anderes bedeuten als bei einer anderen. Und bei ein und derselben *Lehrperson* kann die Bedeutung einer Note von Prüfung zu Prüfung schwanken. Oft liegt in der Zuweisung von Noten zu Leistungen ein beträchtlicher Ermessensspielraum. Selbst in Fächern wie Mathematik oder Physik, die noch am ehesten eine exakte Leistungsmessung erwarten liessen, zeigen Studien, dass bei der Vergabe von Noten subjektive Momente eine nicht unwesentliche Rolle spielen. Offensichtlich erfüllt die Notenskala die drei klassischen Kriterien eines guten Testverfahrens nur beschränkt: *Objektivität* (ein und dieselbe Leistung wird je nach Lehrperson verschieden beurteilt, das heisst Noten sind nicht wirklich objektiv), *Reliabilität* (ein und dieselbe Leistung wird zu verschiedenen Zeitpunkten verschieden bewertet, das heisst Noten sind nur beschränkt verlässlich), *Validität* (die Note, die einer Leistung zugeteilt wird, ist nicht immer ein eindeutiges Zeichen für vorhandenes oder nicht-vorhandenes Wissen, das heisst Noten messen oft nicht, was sie messen sollten).

Die fragliche Qualität von Noten als Messinstrument für schulische Leistungen wird oft damit begründet, dass den Noten verschiedene Funktionen zukommen. Sie haben eine *didaktische* bzw. *Evaluationsfunktion*, das heisst sie sollen nachweisen, ob eine Lektion erfolgreich war oder nicht, ob also der Stoff «angekommen» ist oder nicht. Bei dieser Funktion geht es nicht um die Prüfung der Schülerinnen und Schüler, sondern um die Evaluation des Unterrichts.[10] Noten haben eine *diagnostische* bzw. *Zertifikationsfunktion*, das heisst sie sollen belegen, was ein Schüler oder eine Schülerin weiss. Noten haben eine *Auslese-* bzw. *Selektionsfunktion*, das heisst sie sollen die Schülerinnen und Schüler nach ihrem Leistungsgrad klassifizieren. Noten haben eine *Zeugnis-* bzw. *Berechtigungsfunktion*, das heisst sie weisen einer Lernenden oder einem Lernenden das Recht zu, eine bestimmte weiterführende Schule oder Berufsausbildung zu besuchen (z.B. das Maturitätszeugnis als Berechtigung für den Besuch einer Hochschule).[11] Diese vier Funktionen von Noten setzen eine hohe Objektivität, Reliabilität und Validität voraus.

---

[10] In diesen Kontext gehört auch die Unterscheidung von formativer und summativer Beurteilung. Eine Lehrperson tut gut daran, ihren Unterricht nicht nur am Ende einer Unterrichtseinheit (summativ) zu evaluieren, sondern ihre Vermittlungstätigkeit laufend im Auge zu behalten, das heisst auch den *Prozess* (formativ) zu evaluieren.

[11] In politischer Hinsicht ist dies eine wesentliche Funktion von Schule, die unter einer rein pädagogischen Perspektive leicht übersehen wird: Die Schule *bezeugt*, dass der abgehende Schüler bzw. die abgehende Schülerin über ein bestimmtes Leistungsniveau verfügt, das die aufnehmende Institution als eine Art Ein-

Anders steht es mit den folgenden Funktionen, die den Noten ebenfalls zugewiesen werden. Noten können eine *Motivationsfunktion* haben, das heisst sie werden dafür eingesetzt, um einen Schüler oder eine Schülerin zu weiteren Leistungen zu ermuntern. Sie werden für ihre besondere Anstrengung gelobt und erhalten deshalb eine Note, die bewusst *nicht* objektiv und valide ist. Auch eine schlechte Note kann aus motivationalen Gründen erteilt werden, um zum Beispiel einer allzu unbesorgten Schülerin oder Schüler den Hinweis zu geben, dass sie oder er sich etwas mehr anstrengen soll. Noten kann auch eine *Disziplinierungsfunktion* zukommen, das heisst sie können als Mittel der Bestrafung unerwünschten Verhaltens eingesetzt werden. Das kann die ganze Klasse betreffen: Eine unangekündigte oder eine besonders schwierige Prüfung wird nicht aus diagnostischen oder didaktischen Gründen eingesetzt, sondern als (latente) Strafmassnahme. Aber auch einzelne Schülerinnen und Schüler können mittels Notengebung bestraft werden, indem ihre Leistungen strenger bewertet werden als die Leistungen der anderen. Pädagogisch ist diese Funktion von Noten verpönt, aber alle dürften wir die Erfahrung gemacht haben, dass die Unzuverlässigkeit der Notengebung solchen Missbrauch zulässt.

Die Vielfalt an Funktionen, die Noten haben, führt immer wieder zur Forderung, die Funktionen seien zu entmischen, so dass die Qualität der Notengebung gesteigert werden kann.

Zur Validität der Notengebung gehört auch, dass der Stoff, der in einer Prüfung abgefragt wird, tatsächlich gelehrt worden ist. Noten sind nur dann valide (gültig), wenn sie Leistungen messen, die auf den Unterricht bezogen werden können (und nicht z.B. dank eines Elternteils erbracht werden, der dem Kind den Stoff beigebracht hat). In der Sprache der Testtheorie müssen Prüfungen *kriteriumsorientiert* sein, das heisst anhand des vermittelten Stoffes müssen Kriterien bezeichnet werden können, die erfüllt sein müssen, wenn wir sagen, dieser Schüler bzw. diese Schülerin hat den Stoff dieser Lektion gelernt oder nicht. Kriterien, die anhand des Unterrichtsstoffes formuliert werden, würden eine Leistungsbeurteilung ermöglichen, die rein individuumsbezogen ist. Es liesse sich bei jedem Schüler und bei jeder Schülerin einzeln, das heisst unabhängig von den Leistungen der anderen, feststellen, ob die vorweg definierten Kriterien erfüllt worden sind oder nicht.

Damit wird ein wichtiges Merkmal des Beurteilungsprozesses sichtbar: Menschen können an sich selber gemessen werden, oder sie können mit anderen verglichen werden. Beurteilung hat letztlich immer mit *Vergleichen* zu tun. Wenn wir wissen möchten, worin die Besonderheiten des Menschen liegen, dann vergleichen wir den Menschen mit den Tieren. Wenn wir wissen möchten, worin die Eigenheiten der Männer liegen, dann vergleichen wir sie mit den Frauen. Oder wir vergleichen die Franzosen mit den Deutschen, die Lehrer mit den Ärzten, die Monosexuellen mit den Bisexuellen etc. Wenn wir wissen möchten, worin die Besonderheiten *dieses* Individuums liegen, dann vergleichen wir es mit allen anderen Individuen. Und wenn wir wissen möchten, inwiefern sich dieses Individuum *verändert* hat, dann vergleichen wir es mit sich selbst, das heisst mit Merkmalen, die wir vor einer gewissen Zeit an ihm beobachtet haben. Immer beruht unser Urteil auf einem Vergleich. Anders sind diagnostische Beurteilungen gar nicht möglich. Pointiert ist diese Einsicht von den beiden Amerikanern Clyde Kluckhohn und Henry A. Murray wie folgt auf den Punkt gebracht worden: «Every man is in certain respects like *all* other men, like *some* other men, [and] like *no* other man» (Kluckhohn & Murray 1948, p. 35 – eigene Hervorhebung).

Um die Vergleiche anstellen zu können, die uns erlauben festzustellen, inwiefern der andere wie alle anderen, wie einige andere oder wie kein anderer ist, brauchen wir *Bezugsnormen*. Unterschieden wird in kriteriumsorientierte, individuelle und soziale Bezugsnormen. Kriteriumsorientiert ist eine Bezugsnorm, wenn sie das Kriterium angibt, das erfüllt sein muss, damit ein bestimmtes Urteil

---

trittsbillett respektieren muss. Wir erleben zur Zeit in verschiedenen Bereichen unseres Bildungssystems eine gewisse Entwertung dieser Zeugnisfunktion von Schule, so wenn zum Beispiel Lehrbetriebe Aufnahmeprüfungen oder Assessmentverfahren durchführen. Die Schule erfährt damit ein Stück Entwertung; es wird ihr die Kompetenz der Zertifizierung punktuell abgesprochen.

gefällt (z.B. eine bestimmte Note gegeben) wird. Individuell ist eine Bezugsnorm, wenn sie eine erbrachte Leistung mit früheren Leistungen *derselben* Person vergleicht. So kann eine Schülerin oder ein Schüler ein objektiv gesetztes Kriterium (z.B. 7 von 10 Mathematikaufgaben lösen können) erfüllen (kriteriumsorientierte Bezugsnorm), oder kann eine Leistung erbringen, die besser oder schlechter ist als die Leistungen, die sie oder er bisher bei vergleichbaren Aufgaben erbracht hat (individuelle Bezugsnorm). Was schliesslich die soziale Bezugsnorm anbelangt, so gilt oft die Schulklasse als Vergleichsbasis: Die Leistung von Schüler X wird mit der Leistungsverteilung aller Schülerinnen und Schüler in der Klasse Y verglichen. Diese drei Bezugsnormen werden in der Schule oft undifferenziert miteinander vermischt, was die Qualität der Notengebung beeinträchtigt.

Was die Validität von Noten anbelangt, so liegt eine weitere Grenze der Noten als Messinstrument in ihrem beschränkten Aussagewert. Noten stellen lediglich fest, was der Fall ist (mit mehr oder weniger Genauigkeit und Verlässlichkeit); sie geben keine Auskunft über die *Ursachen* einer erbrachten Leistung. Hat eine Schülerin eine 5.5 erhalten, weil sie in einem Fach begabt ist oder weil sie sich besonders gut auf die Prüfung vorbereitet hat? Hat ein Schüler eine 3 erhalten, weil er ein Faulpelz ist oder weil er nichts begriffen hat? Zwischen Begabung und Anstrengung, diesen beiden wichtigen Ursachen für Leistungen, kann anhand einer Note nicht unterschieden werden. Der Informationswert von Noten ist *gering*. Sie klassifizieren nach gut und schlecht, sagen aber nicht, weshalb jemand zu den Guten oder zu den Schlechten gehört. Wenn von der Schule und den Lehrpersonen erwartet wird, dass sie eine *ganzheitliche* Beurteilung ihrer Schülerinnen und Schüler abgeben, dann sind zusätzliche Informationen notwendig, die einer Prüfung allein nicht entnommen werden können. Hier sieht man, weshalb die Noten und die Notengebung fast permanent unter Kritik stehen. Man versucht heute, den verschiedenen Ansprüchen an die Schülerinnen- und Schülerbeurteilung durch Mischformen von Ziffern- und Verbalzeugnissen gerecht zu werden.

Wenn die Leistungsbeurteilung durch Noten nicht genügt, dann stellen sich einer Lehrperson eine Reihe von Fragen: Wie gewinne ich gültiges (valides) Wissen über das Lernverhalten eines Schülers oder einer Schülerin? Wie kann man verlässliche (reliable) Aussagen über die soziale Kompetenz von Schülerinnen und Schülern machen? Wie kommt man zu *prognostischen* Aussagen, das heisst zu begründeten Vermutungen über die weitere Entwicklung oder Schullaufbahn eines Schülers oder einer Schülerin?

Die Ausweitung des Horizonts über den engen Leistungsbereich hinaus wirft aber auch grundsätzliche Fragen der Schülerinnen- und Schülerbeurteilung auf. Es stellt sich die Frage, was überhaupt Gegenstand der schulischen Beurteilung sein *darf*. Unbestritten ist, dass die Schule ihre (intendierten) Effekte überprüfen soll, das heisst sie soll feststellen, ob der Stoff, den sie gemäss Lehrplan zu vermitteln hat, seitens der Schülerinnen und Schüler tatsächlich angeeignet worden ist. Hat die Schule aber auch das Recht, andere Urteile über Lernende zu fällen, zum Beispiel Urteile über ihre Persönlichkeit, ihren Charakter oder ihre Sozialkompetenz? Die Forderung nach ganzheitlicher Beurteilung ist zumeist so ausgerichtet, dass nicht nur das Arbeits-, Lern- und Leistungsverhalten der Schülerinnen und Schüler beurteilt werden soll, sondern auch Merkmale ihrer Persönlichkeit im weitesten Sinne des Wortes. Ist dies legitim?

Die Frage führt zurück zu den vier Aufgaben von Schule, die wir unterschieden haben (vgl. oben). Wer es ablehnt, dass die Schule auch eine pädagogische Aufgabe hat, wird dazu neigen, die pädagogische Diagnostik auf Leistungsdiagnostik einzuschränken. Wer demgegenüber der Meinung ist, dass Schule und Unterricht die Schülerinnen und Schüler auch in ihrer Persönlichkeit und in ihrem Sozialverhalten fördern soll, der wird den Gegenstand der pädagogischen Diagnostik weiter fassen. Welche Position auch immer eingenommen wird, wir sollten zwischen *Schülerinnen- und Schülerbeurteilung* im weiten Sinne des Wortes und *Leistungsmessung* differenzieren. Die Schülerinnen- und Schülerbeurteilung, die über die blosse Leistungsdiagnostik hinaus geht, muss an die Professionsethik des Lehrerberufs zurückgebunden werden, eine Ethik, der es im Kern um das Wohl des

Kindes und des Jugendlichen geht, die der Lehrerin oder dem Lehrer anvertraut sind. Bei der Schülerinnen- und Schülerbeurteilung kommen sinnvollerweise auch andere Verfahren, die oft unter dem Begriff des *Assessment* zusammengefasst werden, zum Einsatz als Prüfungen (vgl. Gardner 1994; Gipps 1994). Dabei spielen *qualitative* Kriterien eine grössere Rolle als die herkömmlichen quantitativen Kriterien der psychologischen Messtheorie (zu Messtheorie vgl. Abschnitt 1.1). Zu überlegen ist dann, wie den beiden Aufgaben der Schülerinnen- und Schülerbeurteilung und der Leistungsmessung genüge getan werden kann, ohne dass sie sich in die Quere kommen.

Assessmentverfahren werden der pädagogischen Aufgabe der Schule eher gerecht als die Verfahren zur Leistungsmessung. Tests und andere Verfahren der Leistungsmessung (Prüfungen) sind von ihrer Logik her *produktorientiert*, während Assessmentverfahren stärker *prozessorientiert* sind. In der Sprache von formativer und summativer Evaluation ausgedrückt, sind Assessmentverfahren formativ, während die klassischen Testverfahren summativen Charakter haben. Vor allem die Kritik an den in den USA weit verbreiteten Multiple-Choice-Verfahren zeigt, dass die Fixierung an den Produkten, das heisst an den erbrachten Leistungen von Schülerinnen und Schülern, pädagogisch wenig hilfreich ist. Untersuchungen zeigen, dass Lernende, die Prüfungen erfolgreich absolvieren, trotzdem vom Stoff, auf den sich die Prüfungen beziehen, wenig bis gar nichts *verstanden* haben. Wer prüfen möchte, ob seine Schülerinnen und Schüler nicht nur auswendig gelernt, sondern den Stoff auch verstanden haben, kann sich mit standardisierten Verfahren nicht begnügen, sondern muss auch qualitative Methoden einsetzen.

Beurteilung als sozialer Prozess

Wir haben bereits darauf hingewiesen, dass Beurteilung im Kontext von Schule und Unterricht kein Vorgang ist, der zwischen einer Lehrperson und einer einzelnen Schülerin oder einem einzelnen Schüler stattfindet. Unterricht ist ein soziales Geschehen, und das hat Auswirkungen auf den Beurteilungsprozess. Es sei hier kurz auf zwei solche Auswirkungen hingewiesen.

(1) Ein wichtiges Moment der Beurteilungspraxis in der Schule liegt im Feedback, das Beurteilungsformen auf den Unterricht und das Lernverhalten der Schülerinnen und Schüler, aber auch auf das Verhalten der Lehrperson haben. Wenn man Schule und Unterricht nicht im Rahmen eines linearen Modells der Wirkkausalität betrachtet, sondern als *Systeme* interdependenter und rückgekoppelter Prozesse versteht, dann kann die Beurteilungspraxis einer Lehrperson nicht losgelöst vom Gesamtgeschehen im Unterricht gesehen werden. In einer systemischen Betrachtungsweise von Unterricht kann deutlich werden, dass Beurteilungsformen oder allgemeiner gesagt Messsysteme nicht einfach passiv registrieren, was der Fall ist, sondern bei den Betroffenen *Reaktionen* auslösen, die als Anpassungen an das Messmodell verstanden werden können. Im Falle von Schule und Unterricht haben wir es nicht wie bei materiellen Objekten mit einer trägen Materie zu tun, die von den Messfühlern der Qualitätssicherung unberührt bleiben, sondern mit aktiven Subjekten, die auf die Art und Weise, wie sie beurteilt werden, reagieren. Dementsprechend sind Messoperationen in sozialen Systemen nicht mit der physikalischen Temperatur-, Längen- oder Zeitmessung zu vergleichen.

Ein im Unterricht verwendetes Messsystem, wie zum Beispiel ein Multiple-Choice-Test, führt zu Anpassungen der Betroffenen: Die Lehrperson wird ihren Unterricht darauf ausrichten, dass bei den Schülerinnen und Schülern Wissenselemente abgefragt werden, und die Lernenden werden sich darauf einstellen, dass ihr Wissen im Rahmen von vorgegebenen Antwortmöglichkeiten häppchenweise abgefragt wird. Soziale Systeme wie der Unterricht widersprechen damit einer wesentlichen Voraussetzung der klassischen psychometrischen Diagnostik, die von einer klaren Subjekt-Objekt-Trennung ausgeht und mit der Stabilität des gemessenen Objekts rechnet. Die vermeintliche Neutralität der Messung wird unterlaufen durch den systemischen Effekt, den die Tests im Unterricht auslösen. Insofern liegen die Verzerrungen, die den Tests gelegentlich vorgeworfen werden (z.B. den

Intelligenztests im Hinblick auf kulturelle Vorurteile), nicht eigentlich in diesen selbst, sondern in ihren unkontrollierten und kaum kontrollierbaren Nebenwirkungen. Insofern sollte sich jede Lehrperson Gedanken darüber machen, welche möglicherweise unbeabsichtigten Nebeneffekte die Art und Weise ihrer Leistungsbeurteilung auf ihren Unterricht bzw. das Lernverhalten der Schülerinnen und Schüler haben kann.

(2) Ein zweites Moment des systemischen Charakters schulischer Beurteilung liegt stärker auf der individuellen Ebene. Philip Jackson unterscheidet vier Methoden, die einer Lehrperson zur Verfügung stehen, um zu prüfen, inwiefern ihre Lehrtätigkeit erfolgreich ist oder nicht (vgl. Jackson 1986, p. 58ff.). Als erstes kann sie aufgrund nonverbaler Signale (Mimik, Gestik, Gebärden) der Schülerinnen und Schüler auf Verständnis oder Unverständnis schliessen. Lernende verhalten sich auf spezifische Weise, wenn sie verstanden haben oder nicht verstehen. Als zweites kann die Lehrperson für eine Atmosphäre sorgen, in der Verständnis und Unverständnis offen und ehrlich ausgedrückt werden können, wo zum Beispiel Schülerinnen und Schüler, die nicht verstehen, fragen dürfen, ohne dass sie blossgestellt werden, wo die Lehrperson unterbrochen und zur Wiederholung aufgefordert werden darf, wo Fehler gemacht werden dürfen, ohne dass sich dies gleich in negativen Bewertungen niederschlägt, wo Lernende ganz grundsätzlich die Lehrperson um Hilfe angehen können etc. Als drittes kann die Lehrperson selber initiativ werden und ihre Schülerinnen und Schüler fragen, ob sie verstanden haben. Sie kann sich unspezifisch an die Klasse richten, oder sie kann Einzelne aufrufen und prüfen, ob sie verstanden haben oder nicht. Sich an die Klasse zu wenden, ist unverfänglich, weil dabei niemand genötigt wird, sein Unwissen einzugestehen, sich an einzelne Schülerinnen und Schüler zu wenden, kann problematisch sein, weil im Falle des Nichtwissens oder Nichtverstehens die Gefahr der öffentlichen Blamage besteht. Als viertes steht der Lehrperson die Methode der formellen Prüfung zur Verfügung. Prüfungen eignen sich allerdings nicht wirklich, um das Verständnis der Schülerinnen und Schüler *während* des Unterrichtsprozesses zu testen; sie markieren viel mehr den Abschluss einer Unterrichtseinheit und kommen damit in didaktischer Hinsicht bzw. als Instrumente der formativen Evaluation immer schon zu spät.

Die vier bzw. drei Methoden ermöglichen es der Lehrperson, ihr Unwissen hinsichtlich des Verständnisses der Schülerinnen und Schüler und der Wirksamkeit ihrer Unterrichtsmethoden formativ zu überprüfen. Keine der Methoden garantiert jedoch Erfolg. Nonverbales Verhalten kann vorgetäuscht werden. Sich zu Wort melden und seine Verstehensdefizite öffentlich einzugestehen ist nicht jedermanns Sache. Nachfragen, ob die Klasse verstanden hat, kann falsche Antworten auslösen. Einzelne Schülerinnen oder Schüler aufrufen kann die Ungewissheit, ob auch die anderen verstanden haben, nicht ausräumen. Schliesslich stellt die didaktische Frage ein zweischneidiges Schwert dar, denn Lehrpersonen fragen ja nicht, weil sie *selber* nicht wissen, sondern sie möchten wissen, was die Lernenden (nicht) wissen. Insofern hat jede Frage, die eine Lehrperson im Rahmen ihres Lehrauftrags an die Klasse richtet, den Anstrich einer Prüfungsfrage. Lehrerinnen und Lehrer können kaum harmlos fragen. Unweigerlich stehen sie im Verdacht, eine Kontrollfrage zu stellen. Damit aber steht der Versuch der Lehrperson zu prüfen, ob sie verstanden wird, in Gefahr, kontraproduktiv zu sein und Reaktionen auszulösen, die auf die soziale Dynamik der Klasse durchschlagen und den Unterricht bedrohen. Soziale Beziehungen sind auf Vertrauen angewiesen und werden durch ein Klima des Misstrauens destabilisiert. Insofern kann eine Lehrperson, die sich darum bemüht, ihren Unterricht laufend (formativ) zu evaluieren, die Qualität ihres Unterrichts gerade durch ihr Bemühen um Qualität ungewollt untergraben.

Diese systemische Betrachtung des Unterrichts lässt die prinzipielle Frage stellen, wie viel Lehrpersonen über ihre Schülerinnen und Schüler überhaupt wissen sollten. Es kann ein Zuviel an Wissen geben bzw. ein Verhalten, das die Lernenden irritiert, weil die Lehrperson gewissermassen zu viel über sie erfahren möchte, ein Zuviel, das sie zur Verweigerung motiviert und damit die Qualität des Unterrichts massiv beeinträchtigt.

Überblick über das Skript

Die pädagogisch-psychologische Diagnostik hat nach gut 100 Jahren ein anspruchsvolles Niveau erreicht. Es kann nicht Ziel dieses Skripts sein, die weit verzweigte Thematik der pädagogischen Urteilsbildung umfassend darzustellen. Vielmehr geht es um eine grundlegende Einführung, die allerdings trotz allem nicht anspruchslos sein wird. Unser Hauptanliegen ist es, die Logik diagnostischer Verfahren vorzustellen, eine Logik, die nicht an einzelne Methoden gebunden ist, sondern eine bestimmte Haltung gegenüber dem Prozess der Urteilsbildung zum Ausdruck bringt. So ist es bei jedem Urteilsprozess wichtig, auf Kriterien zu achten, die das Urteil objektiv, valide und reliabel machen. Nur wenn wir diesen und anderen Kriterien gerecht werden, können wir sagen, dass wir unsere diagnostischen Urteile auf professionellem Niveau fällen.

Im anschliessenden Kapitel 1 werden die *methodischen Grundlagen* der pädagogischen Diagnostik wie Messen und Messtheorie (Abschnitt 1.1), Skalentypen (Abschnitt 1.2), die Grundzüge der klassischen Testtheorie (Abschnitt 1.3), die damit verbundenen Gütekriterien (Abschnitt 1.4) sowie die Anwendung und Interpretation von Normen (Abschnitt 1.5) besprochen. Die hier vermittelten methodischen Grundlagen sind für das Verständnis der darauf folgenden Vertiefungskapitel zentral.

Im Kapitel 2 steht die *schulische Leistungsdiagnostik* im Zentrum: Es wird der Leistungsbegriff eingeführt (Abschnitt 2.2), die Diagnosekompetenz von Lehrpersonen (Abschnitt 2.3) diskutiert und es werden nicht-standardisierte Verfahren zur schulischen Leistungsdiagnostik wie mündliche Prüfungen, Aufsatzprüfungen, Selbstbeurteilung und das Portfolio vorgestellt (Abschnitt 2.4). Anschliessend wird das Thema der Notengebung eingehend behandelt, indem zunächst auf deren Kritik und anschliessend auf Verbesserungsmöglichkeiten hingewiesen wird (Abschnitt 2.5). Danach wird auf standardisierte Schulleistungstests eingegangen und als ausgewähltes Verfahren das Klassencockpit näher betrachtet (Abschnitt 2.6). Als ein Beispiel aktueller Entwicklungen wird sodann die Leistungsbeurteilung durch ausserschulische Instanzen anhand des Basis-Check und des Multi-Check vorgestellt (Abschnitt 2.7). In einem letzten Punkt dieses Kapitels werden schliesslich Grenzen der schulischen Leistungsbeurteilung aufgezeigt (Abschnitt 2.8).

Das Kapitel 3 steht im Zeichen der *Begabungsdiagnostik*: Anhand der Hochbegabung werden unterschiedliche Verfahren der Intelligenz- und Begabungsdiagnostik (Abschnitt 3.2) und mit dem HAWIK-III ein ausgewählter Intelligenztest vorgestellt (Abschnitt 3.3). Es wird auf die Anwendung von Intelligenztests bei Lernstörungen wie Legasthenie oder Dyskalkulie (Abschnitt 3.4) eingegangen und der Einsatz von Intelligenztests, unter anderem in Bezug auf die Diagnostik der Hochbegabung, kritisch reflektiert (Abschnitt 3.5).

Im Kapitel 4 steht das diagnostische Instrument der *Unterrichtsbeobachtung* im Zentrum: Es wird aufgezeigt, was unter Verhaltensbeobachtung zu verstehen ist (Abschnitt 4.2), welche Formen der Beobachtung unterschieden (Abschnitt 4.3) und wie Beobachtungen protokolliert werden (Abschnitt 4.4). Schliesslich wird aufgezeigt, welche Beobachtungssysteme, speziell auch für die Unterrichtsbeobachtung, zur Anwendung vorliegen (Abschnitte 4.5 und 4.6) und wo allenfalls deren Grenzen zu orten sind (Abschnitt 4.7).

1 Methodische Grundlagen der pädagogischen Diagnostik

Nachdem in der Einleitung dieses Skripts der Gegenstand der pädagogischen Diagnostik umrissen und mögliche Handlungsfelder aufgezeigt wurden, sollen im folgenden Kapitel die nötigen methodischen Grundlagen erarbeitet werden. Es soll darauf hingewiesen werden, dass diese Grundlagen hier nur einführend und überblicksartig behandelt werden können; jedes der nachfolgend besprochenen Themen füllt alleine ganze Lehrbücher.[12] Die nun betrachteten Grundlagen des Messens oder der klassischen Testtheorie sind dabei nicht spezifisch für die pädagogische Diagnostik, sondern bilden auch die methodische Basis in anderen empirisch-sozialwissenschaftlichen Disziplinen wie etwa der Psychologie oder der Soziologie.

Seit jeher ist Diagnostik eine zentrale Aufgabe in pädagogischen Handlungsfeldern. Die Frage ist jedoch, inwiefern dieses diagnostische Vorgehen (mit oder ohne diagnostische Instrumente) als ein Vorgehen verstanden wird, in dem unter Beachtung *wissenschaftlicher Gütekriterien* beobachtet und befragt wird, sowie die Ergebnisse sinnvoll interpretiert und mitgeteilt werden (vgl. Ingenkamp 1997, p. 11). Der Unterschied zwischen wissenschaftlich und vorwissenschaftlich soll dabei nicht als prinzipiell, sondern eher als graduell eingestuft werden. Diese Einführung in methodische Grundlagen soll der Leserin und dem Leser aufzeigen, was es bedeuten könnte, Kriterien, die auf die «Richtigkeit» oder «Falschheit» einer Aussage oder einer Folgerung im diagnostischen Prozess hinweisen, zu *überprüfen*. «Methoden, die den Kriterien nicht standhalten, sollten verbessert oder – wenn sich dies nicht erreichen lässt – aus dem Kanon der diagnostischen Verfahren ausgeschieden werden» (Lukesch 1998, p. 37).

Wir bewegen uns im Bereich der methodischen Grundlagen in einem Grenzbereich zwischen Forschung/Wissenschaft und praktischem Handeln in pädagogischen Feldern. Eine klare Trennlinie zwischen Theorie und Praxis herauszuarbeiten wäre weder möglich noch sinnvoll. Zur Veranschaulichung der dargestellten methodischen Grundbegriffe und Konzepte sollen immer wieder Bezüge zum schulischen Alltag gegeben werden, welche aufzeigen, dass solche Kenntnisse nicht nur hilfreich, sondern notwendig sind, wenn es um Bewertung, Beurteilung, Diagnose oder Evaluation geht.

1.1 Messen und Messtheorie

Wer sich mit Diagnostik auseinandersetzt, kommt nicht umhin, sich einige grundsätzliche Gedanken zum Messen und zur Messtheorie zu machen.[13] Dies, da es im Zusammenhang mit Diagnostik oft ums Messen – meist in Form des Testens – geht. Wenn eine Lehrperson Schülerinnen und Schülern für ihre Leistungen, z.B. für einen Aufsatz, eine Note erteilen, dann nimmt die Lehrperson eine Messung vor. Das heisst, die Lehrperson gibt ein Schätzurteil darüber ab, wie sie die *Ausprägung* bestimmter Merkmale bei diesem Aufsatz beurteilt (z.B. Merkmale wie Aufbau/Struktur, formale oder inhaltliche Kriterien) und ordnet diesem Ausprägungsgrad eine Ziffer – in diesem Falle eine Note von 1 bis 6 – zu. Messen bedeutet also vereinfacht gesagt die *Zuordnung von Zahlen zu Beobachtungen.*

Der Zweck einer Messung ist es allgemein, *Informationen über Merkmale* von Objekten, Organismen oder Ereignissen zu erhalten (vgl. Ingenkamp 1997, p. 28). Gemessen werden jeweils nur At-

---

[12] Für detailliertere Ausführungen sei auf die einschlägige Literatur zur Pädagogischen Diagnostik verwiesen. Einen praxisorientierten Überblick gibt etwa Ingenkamp (1997). Theoretisch und methodisch ausführlicher sind Tent und Stelzl (1993) oder Lukesch (1998).

[13] Für eine ausführlichere Einführung in die Messtheorie wird auf Orth (1974) verwiesen.

tribute oder Merkmale von Objekten, nicht die Objekte an und für sich. Also beispielsweise die Länge, die Höhe oder das Gewicht eines Stuhls, nicht der Stuhl als solcher.

Ein wichtiges Ziel des Messens, d.h. der Quantifizierung von Merkmalen, ist die *Exaktheit der Beschreibung* des Objekts sowie die *Vergleichbarkeit* dieses Objekts mit anderen, ähnlichen Objekten. Im oben genannten Fall des Aufsatzes könnte man sich auch vorstellen, dass die Lehrperson den Aufsatz verbal genau beschreibt. Ob allerdings andere Beurteiler den Aufsatz verbal ebenso beschreiben würden oder ob sie aus der Beschreibung dieselben Folgerungen schliessen würden (z.B. dass das ein sehr guter Aufsatz war), ist fraglich. Deshalb vereinfacht die Zensurengebung die Vergleichbarkeit der Leistungen der Schüler und Schülerinnen untereinander. Das Beispiel zeigt aber auch, dass Messwerte immer auch eine *Vergröberung* und *Vereinfachung* einer differenzierten Beobachtung darstellen können.

Eine gängige Definition des Messens gibt Orth (1974):

> «Messen erfolgt durch eine Zuordnung von numerischen Grössen (Zahlen, Vektoren) zu Dingen, die Träger der zu messenden Eigenschaft sind. Messen beruht auf einer homomorphen Abbildung eines empirischen Relativs in ein numerisches Relativ bzw. auf einer Repräsentation eines empirischen Relativs durch ein numerisches Relativ. Die Existenz einer derartigen homomorphen Abbildung ist das Kriterium dafür, ob eine Zuordnung von Zahlen zu Dingen als ‹Messen› zu betrachten ist, d.h. ob eine Eigenschaft messbar ist» (Orth 1974, p. 18).

Ein *empirisches Relativ* liegt dann vor, wenn eine Menge empirischer Dinge mit bestimmten Relationen vorliegt, z.B. Schülerinnen und Schüler mit der Eigenschaft «Intelligenz» und der Ordnungsrelation «intelligenter als» oder die Anzahl richtiger Wörter im Diktat als «Menge empirischer Dinge» und die Relation «mehr richtige als». Ausformuliert hiesse das etwa, Schüler A hat im Diktat mehr richtige Wörter als Schüler B und dieser wiederum mehr richtige Wörter als Schüler C (vgl. Ingenkamp 1997, p. 28). Aus dieser Darstellung wird ersichtlich, dass mit der Relation «mehr richtige als» die Rangfolge der Schüler (A = 1. Rang, B = 2. Rang etc.) dargestellt werden kann und somit die empirische Relation der drei Schüler untereinander (in Bezug auf die richtigen Wörter) durch ein *numerisches Relativ* – in diesem Falle die Rangfolge – sinnvoll und eindeutig (*homomorph*[14]) abgebildet werden kann. Abbildung 1.1 soll diesen Zusammenhang von empirischem und numerischem Relativ grafisch verdeutlichen.

---

[14] Die Abbildung in einem Messvorgang ist dann eindeutig (homomorph), wenn jeder Angabe in den Ursprungsaufzeichnungen genau ein Zahlwert zugeordnet wird. Ein System A ist in einem System B homomorph abgebildet, wenn a) von jedem Bestandteil von A auf genau einen Bestandteil von B (Merkmal => Messwert) und b) von jeder Relation zwischen Bestandteilen von A auf eine Relation zwischen Bestandteilen von B (z.B. Ausprägungen des Merkmals sind geordnet => Messwerte bilden Rangfolgen) geschlossen werden kann, jedoch nicht umgekehrt. Ein Beispiel hierfür wäre Gebäude => Grundriss / Grundriss ≠>Gebäude (viele Gebäude mit demselben Grundriss sind nämlich denkbar!). Oder ein anderes Beispiel wären Fahrzeiten von Skirennfahrerinnen => Rangplätze / Rangplätze von Skirennfahrerinnen ≠> Fahrzeiten. Wäre der Rückschluss vom Messwert auf das empirische Relativ ebenfalls eindeutig möglich, wäre die Abbildung ein-eindeutig (= *isomorph*).

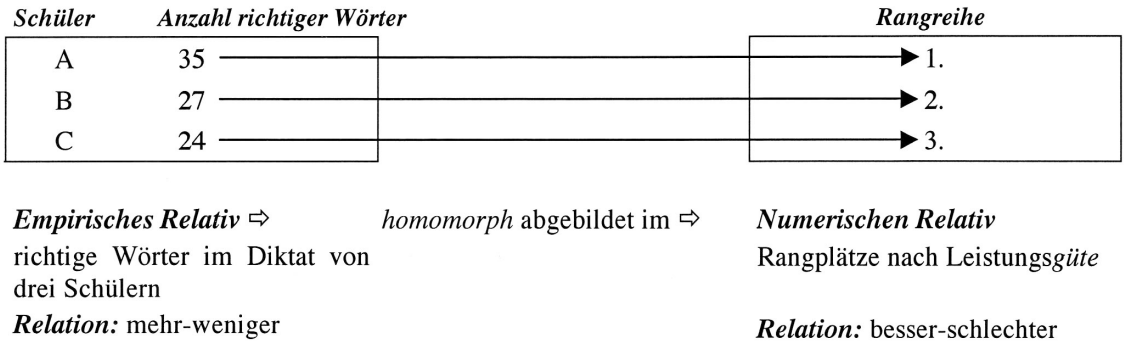

Abbildung 1.1: Empirisches und numerisches Relativ (angelehnt an: Ingenkamp 1997, p. 29)

In der Messtheorie spricht man von der *Repräsentativität der Messung,* wenn es um die Frage nach der Qualität der Abbildung geht: «Das empirische Relativ (das Merkmalsgefüge) soll so im numerischen Relativ abgebildet werden, dass die Beziehungen des empirischen Relativs auch in den Beziehungen des numerischen Relativs auftreten» (Fisseni 1997, p. 28).

Weniger eindeutig wird die Zuordnung des empirischen zum numerischen Relativ dann, wenn etwa die Qualität eines Vortrags oder eines Aufsatzes anhand einer Zahlenskala repräsentiert werden soll, da hier bei einer Bewertung nicht nur ein offensichtliches Merkmal (wie z.B. die Anzahl richtiger Wörter) miteinfliesst. Es geht auch um die Fragen des Stils, der Rhetorik, des Aufbaus oder der Gedankenführung etc. Es müsste also erstens vorerst festgelegt werden, welche Kriterien überhaupt bewertet und zweitens, wie diese numerisch abgebildet werden sollen, «Wir müssten erst eine Reihe von Schätzskalen [...] zwischenschalten und hätten dann auch erst für verschiedene Merkmale, aber nicht für den Vortrag insgesamt eine numerische Repräsentation» (Ingenkamp 1997, p. 29).

1.1.1 Messung psychologischer Merkmale und Eigenschaften

Wenn es um die Messung komplexer psychologisch oder pädagogisch relevanter Merkmale wie etwa Schulerfolg oder Eigenschaften (Dispositionen, «traits») wie etwa Intelligenz, Leistungsmotivation, Konzentration oder Aggressivität geht, stellt sich wie im obigen Fall die Frage, wie diese erfasst und gemessen werden können. Wie aufgezeigt werden wird, ist diese Frage nicht einfach zu beantworten, da solche Merkmale, im Gegensatz etwa zu der Anzahl Fehler in einem Diktat, *nicht direkt beobachtbar* sind. Auch ist es nicht ohne weiteres klar, was mit solchen eher abstrakten Begriffen wie z.B. «Leistungsmotivation» oder «Intelligenz» gemeint sein soll. Vielleicht verstehen unterschiedliche Personen intuitiv etwas anderes darunter?

Fisseni betrachtet dies als eines der Grundprobleme bei der Beschäftigung mit Diagnostik:

> «In der Psychologie sind die *Variabeln* nicht vorgegeben, sondern müssen ‹*konstruiert*› werden. (In der Physik sind manche Grössen vorgegeben, wenigstens für den ersten Blick, etwa Länge, Breite, Höhe). In der Psychologie ist kein ‹Gegenstand an sich› gegeben, weder ‹Intelligenz› oder ‹Konzentration› noch ‹Stimmung› oder ‹Zufriedenheit›. Der messende Psychologe muss seine Gegenstände immer wieder neu definieren – ‹abgrenzen›» (Fisseni 1997, p. 23f.; Hervorhebungen im Original).

Man könnte sich vorstellen, dass Intelligenz z.B. über das Lösen einer kreativen Problemlöseaufgabe (z.B. der Turm von Hanoi[15]), über die Art, wie praktische Probleme im Alltag gelöst werden, oder – wie es traditionell eher gemacht wird – über verschiedene Tests zum Gedächtnis, zu schlussfolgerndem Denken, zu sprachlichen und mathematisch-räumlichen Fähigkeiten erfasst wird.

[15] Der Turm von Hanoi ist ein klassisches Knobelspiel, bei dem es darum geht, einen Turm aus drei unterschiedlich grossen Scheiben von einem Stab auf den andern zu bewegen. Dabei darf man allerdings nur immer eine Scheibe bewegen, und man darf eine grössere nie auf eine kleinere Scheibe legen (vgl. Mathematische Basteleien: www.mathematische-basteleien.de/hanoi.htm).

Denkbar wäre, dass auch musische sowie kreative oder soziale Fähigkeiten erfasst und gemessen werden.

Es wird ersichtlich, dass sich hinter dem abstrakten *theoretischen Konstrukt* Intelligenz[16], numerisch erfasst z.B. durch den Intelligenzquotienten (IQ), eine Menge von möglichen theoretischen Annahmen (theoretische Modelle) «verstecken» können bzw. solche theoretischen Konstrukte theoretische Implikationen enthalten.

> «Ein [psychologisches] Merkmal ist in jedem Fall die begriffliche Fassung eines Aspekts, in dem sich Individuen voneinander und zu verschiedenen Zeiten unterscheiden können. Das begriffliche Abstraktionsniveau kann dabei zwischen der Wiedergabe eines unmittelbar beobachteten Verhaltens, z.B. Sich-melden im Deutschunterricht am Freitag letzter Woche, und erschlossenen theoretischen Konstrukten, wie ‹Lerneifer›, ‹Selbstbewusstsein›, und ‹Geltungssucht›, variieren. Den unterschiedlichen Aussageebenen entsprechen Unterschiede im theoretischen Status der Aussage. Je höher das Abstraktionsniveau, desto mehr Implikationen sind darin enthalten» (Tent & Stelzl 1993, p. 27).

Konstrukte sind allgemein gesagt sprachliche Kürzel, die inhaltlich definierte Bereiche von nicht direkt beobachtbarem oder beschreibbarem Verhalten von Individuen zusammenfassen. *Latente Merkmale* oder theoretisch angenommene psychologische Konstrukte[17] wie etwa Eigenschaften (Dispositionen, «traits») sind nicht unmittelbar gegeben oder direkt ersichtlich, sondern es muss festgelegt bzw. erforscht werden, über welche beobachteten Merkmale oder über welches beobachtbare Verhalten (*manifeste Merkmale* oder *Indikatoren*) diese erschlossen werden können (vgl. Abbildung 1.2).

Latentes Merkmal (angenommenes psychologisches Konstrukt, z.B. „Lerneifer“) → wird erschlossen aus → ***manifestem Merkmal*** (tatsächlich beobachtbare Verhaltensweise als Indikator, z.B. „Sich-Melden-im-Deutschunterricht“)

Abbildung 1.2: Zusammenhang von latenten und manifesten Merkmalen in der pädagogisch-psychologischen Diagnostik

Dieser Zusammenhang zwischen latentem und manifestem (= beobachtbarem) Merkmal soll an folgendem Beispiel illustriert werden: Nehmen wir an, dass ein Kind im Unterricht häufig andere stört, schnell wütend wird, Kameraden ärgert und schlägt und der Lehrperson freche Antworten gibt. Sehr wahrscheinlich versuchen wir, diese Situationskonstellation in einer bestimmten Weise zu verarbeiten und bestimmte Schlüsse daraus zu ziehen, wie etwa, dass dieses Kind «aggressiv» sei. Wir machen also einen Rückschluss auf eine *Eigenschaft* (Disposition, «trait») des Kindes. Die Eigenschaft ist nicht unmittelbar und sinnlich wahrnehmbar (*phänomenal*) gegeben. Dieser Schluss auf psychische Dispositionen ist ein kreativ-konstruktiver gedanklicher Akt:

---

[16] Auf das Thema «Intelligenz» und Intelligenzdiagnostik wird im Kapitel 3 zur Begabungsdiagnostik ausführlich eingegangen.

[17] Zur Abgrenzung der Begriffe: Mit dem Begriff «Konstrukt« können nicht nur psychologische, sondern auch andere abstrakte, nicht direkt beobachtbare oder beschreibbare Dinge oder Erscheinungen gemeint sein, welche in einem gedanklichen Akt *konstruiert* werden müssen. Beispiele wären etwa «Gesundheit», «Weltfrieden» oder «Kaufkraft». Der Begriff «Konstrukt» ist in diesem Sinne ein allgemeinerer oder übergeordneter Begriff. Die Begriffe «latente» und «manifeste Merkmale» hingegen stammen aus der Persönlichkeitspsychologie und werden in diesem Sinne enger, nur zur Beschreibung des Zusammenhanges von persönlichkeits- oder eigenschaftsrelevanten Merkmalen verwendet (Aggressivität, Leistungsmotivation, Introversion, Extraversion, Intelligenz, Ängstlichkeit etc.).

«Wir projizieren das beobachtete Verhalten sozusagen ‹nach innen›, indem wir uns vorstellen, es sei die Manifestation einer (momentanen oder dauerhaften) psychischen (respektive physischen) Disposition unseres Verhaltensträgers [...], bestimmte Situationskonstellationen in einer bestimmten Weise zu verarbeiten und mit einem entsprechenden Verhalten zu beantworten» (Beck & Krapp 2001, p. 56).

Angefügt werden soll, dass wir im Alltag solche Rückschlüsse auf Eigenschaften relativ schnell und beinahe automatisch vollziehen (beobachten Sie sich einmal!) und ziemlich rasch bereit sind, einer Verhaltensweise (z.B. Schweigen) eine psychische Eigenschaft zuzuschreiben (z.B. Schüchternheit). In professionell-pädagogischen Kontexten sollte man sich jedoch bewusst sein, dass solche Zuschreibungen ev. auch einmal vorschnell getroffen wurden oder dass zu wenig Information, z.B. auch über den Kontext, miteinbezogen worden ist. Eine solche gedankliche Hypothese sollte im Sinne einer Arbeitshypothese auch revidierbar sein, d.h. wir müssen uns auch immer wieder von neuem hinterfragen, ob sie noch Gültigkeit besitzt. In der Sozialpsychologie konnte gezeigt werden, dass Menschen eine Tendenz haben, Verhaltensweisen *anderer* Personen auf *Eigenschaften* zu attribuieren (= zuschreiben), statt sie z.B. der Situation zuzuschreiben.[18] Es sollte zumindest in Erwägung gezogen werden, dass eine Schülerin, welche bei einer Begegnung nicht viel spricht, ev. gerade starke Bauchschmerzen hat oder gedanklich beim Streit mit ihrer besten Freundin weilt.

Zurück zum Beispiel der Intelligenz: Wenn angenommen wird, dass theoretisch (also je nach Stand der Theoriebildung oder je nach Theorie, auf die wir uns stützen) auch die «soziale Intelligenz» als zur Intelligenz gehörig angesehen wird, dann fragt sich nun in einem ersten Schritt, was unter sozialer Intelligenz gemeint ist, resp. welche Definitionsvorschläge wir dazu vorliegend haben. Gardner (1983) etwa hat eine Theorie der Multiplen Intelligenzen vorgeschlagen, in der einer der sieben Bereiche der Intelligenz, die er vorschlägt, die *Interpersonelle Intelligenz* ist. «Diese Form der Intelligenz, häufig auch als soziale Intelligenz bezeichnet, hat mit der Fähigkeit zu tun, subtile Hinweise und Anspielungen in unserer komplexen sozialen Umwelt (Elternhaus, Freundeskreis, Schule, Verein oder Nachbarschaft) aufzugreifen und zu verstehen» (Gage & Berliner 1996, p. 79).

Wenn wir nun aus einem theoretischen Modell abgeleitet eine Definition von sozialer Intelligenz haben, wäre die zweite Frage, wie diese *operationalisiert* (messbar gemacht) werden kann. Konkret wird nun also versucht, die Fähigkeit, «subtile Hinweise und Anspielungen in unserer komplexen sozialen Umwelt aufzugreifen und zu verstehen», messbar zu machen. Vielleicht würde man den untersuchten Personen soziale Situationen (z.B. dargestellt in Bildergeschichten) präsentieren und man würde ermitteln, wie gut sie diese verstehen und interpretieren können. Oder man könnte versuchen zu messen, wie gut jemand mimische Gefühlsäusserungen bei einem Gegenüber decodieren, d.h. wahrnehmen und richtig interpretieren, kann. Dies wäre ein Zeichen für seine Empathie (Einfühlungsvermögen).

Soziale Intelligenz könnte aber auch über andere Indikatoren wie die Anzahl der Freundinnen und Freunde oder die Beliebtheit bei Klassenkameradinnen und -kameraden erschlossen werden. Dies wäre gerechtfertigt, da Untersuchungen gezeigt haben, dass Personen mit speziellen sozialen Kompetenzen (z.B. Konfliktlösestrategien, kommunikative Fähigkeiten oder Fähigkeit zur Kontrolle aggressiver Handlungsimpulse) tendenziell eine grössere Anzahl von Freundinnen und Freunden und eine grössere Popularität aufweisen als Schülerinnen und Schüler mit unangemessenem Sozialverhalten (vgl. Wild, Hofer & Pekrun 2001, p. 259f.).

Es wird ersichtlich, dass die Antwort auf die *Frage nach der Operationalisierung psychischer Merkmale* verschieden ausfallen kann. Darüber hinaus zeigt sich auch, dass sich gerade hinter Kon-

---

[18] In der Attributionsforschung spricht man hier vom «fundamentalen Attributionsfehler». Interessant ist auch zu wissen, dass dieselben Personen, wenn es um die Erklärung ihres *eigenen* Verhaltens geht, viel eher die *Situation als Auslöser* für ihr Verhalten (eben z.B. zu schweigen) verantwortlich machen als eine Eigenschaft bei ihnen (vgl. dazu auch APP-Modul «Sozialpsychologie des Unterrichts»).

strukten auf einer hohen theoretischen Ebene (wie z.B. «Intelligenz») wiederum andere theoretische Begriffe oder Konstrukte verbergen können (wie z.B. «Empathie» oder «kommunikative Fähigkeiten»), welche wiederum (theoretisch sowie auch operational) definiert werden müssen. Es wird also deutlich, dass es für ein und dasselbe theoretische Konstrukt viele «operationale Definitionen» geben kann (vgl. Beck & Krapp 2001, p. 58).[19]

1.2 Skalentypen

Zurück in die schulische Praxis: Wenn man sich noch einmal das Beispiel der drei Schüler und der Anzahl richtiger Wörter im Diktat vergegenwärtigt (35, 27 und 24 richtige Wörter), dann wird ersichtlich, dass mit der Festlegung der Rangfolge der drei Schüler zwar *eine* Information gegeben ist, diese aber in Hinblick auf gewisse Fragestellungen noch recht undifferenziert ist, weil sie keine Information über die *Intervalle* enthält, weil die Abstände in der gemessenen Leistung (z.B. in Sekunden oder Millisekunden in einem Sportwettkampf) zwischen den Rangplätzen jeweils gleich behandelt werden, auch wenn sie absolut gesehen überhaupt nicht gleich gross ausgefallen sind. Aber auch die Unterschiede zwischen den Anzahlen richtig geschriebener Wörter (35-27=8, 27-24=3 und 35-24=11) sagen wenig über die *Grösse der Unterschiede* in der *Rechtschreibe-Fähigkeit* der drei Schüler aus. Wohl lässt sich aber vermuten, dass Schüler 35 über die «grössere» Rechtschreibe-Fähigkeit verfügt als Schüler 27 oder Schüler 24.

Das Beispiel zeigt, dass wir uns bei Messungen auf *unterschiedlichen Exaktheitsniveaus* bewegen können. Meist werden dabei *vier Niveaus* unterschieden, welche durch *vier Skalen* gekennzeichnet werden. Ein Ziel ist es, Daten auf einem möglichst hohen Niveau darzustellen. Die Anwendung eines Skalentyps hängt aber massgeblich von den Daten, d.h. deren Beschaffenheit (Qualität) ab und ist an gewisse Voraussetzungen gebunden. Nur wenn die Voraussetzungen eines Skalentyps erfüllt sind, kann auf diesem Niveau mit den Daten operiert werden bzw. können gewisse Rechenoperationen durchgeführt werden.[20] Dies wird im Folgenden dargestellt.

1.2.1 Nominalskala

Das «unterste» Messniveau oder der tiefste Skalentyp ist die *Nominalskala*, sie ist an die geringsten Voraussetzungen der Datenqualität geknüpft. Bei «Messungen»[21] auf diesem Skalentyp geht es nur um *qualitative Gleichartigkeit* von Merkmalen, d.h. um die Frage, ob verschiedene Objekte sich bezüglich eines Merkmals unterscheiden oder nicht. Beispiele hierfür könnte die Einteilung von Schülerinnen und Schülern nach ihrer Konfessionszugehörigkeit (katholisch, reformiert, buddhistisch, muslimisch etc.), nach Geschlecht, Nationalität, Beruf der Mutter oder nach ihrer Augenfarbe sein. Diese Kategorisierung geschieht v.a. um Voraussetzungen für das angemessene *Zählen* (Häufigkeitszählungen) zu schaffen. Wichtig ist, dass eine eindeutige Zuordnung zu einer Kategorie

---

[19] Es ist hier v.a. Aufgabe der Wissenschaft und der Theoriebildung, erstens geeignete theoretische Modelle vorzuschlagen, zweitens Möglichkeiten zu finden, diese an die Realität anzubinden – d.h. zu operationalisieren – und diese drittens schliesslich auch empirisch zu überprüfen und weiterzuentwickeln *(Validierung)*.

[20] In der Messtheorie ist dies die *Bedeutsamkeit* einer Messung, d.h. die Frage, welche mathematisch/statistischen Operationen für die einzelnen Skalen erlaubt sind nebst der *Repräsentativität,* auf die schon kurz hingewiesen wurde (vgl. Abschnitt 1.1), wäre die *Eindeutigkeit* das dritte wichtige Stichwort, um eine Messung zu charakterisieren. Eindeutigkeit betrifft nach Fisseni (1997) die Transformation, die im numerischen Relativ mathematisch möglich ist, ohne die Relationen des numerischen Relativs zu verfälschen (vgl. Fisseni 1997, p. 28).

[21] Nach Ansicht einiger Autoren (u.a. Jones 1971; Kriz 1973) sollte hier noch nicht von einem Messvorgang gesprochen werden, da eine Zuordnung von Zahlen nur eine «namengebende Aufgabe» hat und allen Zahlen somit lediglich Markierungsqualität zukommt (vgl. Kleber 1992, p. 161).

möglich ist – andernfalls muss die Kategorie verändert oder erweitert werden (vgl. Ingenkamp 1997, p. 29f.).

Da bei einer Nominalskala unterschiedliche Zahlen nicht für ein «Mehr» oder «Weniger», «Grösser» oder «Kleiner» stehen, sind Berechnungen wie Addition oder Multiplikation nicht gestattet bzw. sinnlos.[22] Das Mass der zentralen Tendenz (d.h. der mittlere Wert) ist hier der *Modalwert*, d.h. der am häufigsten auftretende Wert. Wenn 4 Kinder blaue Augen, 7 braune und 2 grüne Augen haben, dann wäre der Modalwert 7.

1.2.2 Ordinal- oder Rangskala

Der zweite Skalentyp ist die *Rang- oder Ordinalskala*. Wenn man sich wiederum das Beispiel mit den richtigen Wörtern im Diktat vergegenwärtigt, dann wird ersichtlich, dass die Schüler anhand der Anzahl richtig gelöster Wörter in eine *Rangordnung* gebracht worden sind (vgl. Abbildung 1.1). Wie schon angetönt sagt diese *Reihenfolge* jedoch nichts über die Abstände der drei Schüler untereinander aus, denn die empirische Realität (35, 27 und 24 richtige Wörter) wird durch die Rangplätze 1, 2 und 3 nicht gleichwertig abgebildet. Es ist deshalb auch nicht sinnvoll, Rangplätze zu dividieren, zu multiplizieren oder mit ihnen arithmetische Mittelwerte zu berechnen. Ebenso wenig kann bei dieser Skala etwas über Verhältnisse ausgesagt werden, wie etwa, dass Schüler A (1. Rang) dreimal so viele richtige Wörter im Diktat hatte wie Schüler C (3. Rang). Wenn man es durchrechnet, sieht man, dass A 1,458 Mal mehr richtige Wörter hat als C.

Das Mass der zentralen Tendenz ist bei einer Ordinal- oder Rangskala der *Median*. Es ist derjenige Wert, der die Stichprobenbeobachtungen in zwei gleich grosse Hälften (50% liegen darunter, 50% darüber) aufteilt. In unserem Beispiel wäre der Median die Leistung des Schülers B.

Wie Ingenkamp (1997, p. 31) betont, ist auch die *Zensurenskala* im Allgemeinen eine Ordinalskala, denn die Abstände zwischen den einzelnen Notenstufen bilden meist *nicht* gleiche Abstände zwischen realen Leistungen ab. Wir erfahren zwar, dass Schüler A in dieser Klasse besser ist als B und dieser wiederum besser als Schüler C. «Aber niemand kann behaupten, dass der Leistungsunterschied zwischen den Zensuren 1 und 2 gleich gross wäre wie der zwischen den Zensuren 3 und 4 oder der zwischen den Zensuren 5 und 6. Diese Anforderungen an die Gleichheit der Skalenabstände (*Äquidistanz*) sind in der Realität der Verhaltensmerkmale nicht nachzuweisen» (ebd.). Deshalb dürften – sofern korrekt vorgegangen würde – eigentlich keine arithmetischen Mittelwerte, d.h. keine Durchschnittswerte berechnet werden, auch wenn dies viele Lehrpersonen dennoch tun und dies in den Prüfungsordnungen immer wieder gefordert wird.[23]

1.2.3 Intervallskala

Wie gezeigt wurde, liegen auf den beiden untersten Messniveaus keine gleichen Skalenabstände vor. Diese Voraussetzung ist erst auf dem *Intervallskalenniveau* gegeben, d.h. erst hier treten bei gleichen Leistungsunterschieden an unterschiedlichen Stellen der Skala auch gleiche numerische Differenzen der Zahlenwerte auf. Somit bildet diese Skala gleich grosse Differenzen im empirischen Relativ «richtig» ab (vgl. Kleber 1992, p. 162).

---

[22] Man stelle sich als nicht ganz ernst gemeintes Gedankenexperiment vor, dass die Kategorisierung wie folgt aussähe: katholisch = 1, reformiert = 2 und muslimisch = 3. Wenn man nun je ein reformiertes und ein katholisches Kind ermittelt hätte, und diese «addieren» würde, käme man dann auf ein muslimisches Kind? Oder ergäbe der Durchschnitt (Mittelwert) von einem katholischen und einem muslimischen Kind ein reformiertes Kind?

[23] Vgl. zum Thema Notengebung auch Abschnitt 2.5.

Ein Beispiel für eine Intervallskala wäre die Einteilung in Grad Celsius. Zwischen 25° und 50° Celsius besteht beispielsweise dieselbe Energiedifferenz wie zwischen 50° und 75° Celsius. Was jedoch nicht zulässig ist, ist die Behauptung, dass 100° Celsius doppelt so heiss sei wie 50° Celsius, da der Nullpunkt 0° Celsius willkürlich festgelegt wurde – der natürliche Nullpunkt wäre nämlich bei –273° Celsius. Wenn wir den natürlichen Nullpunkt verwenden (0 Kelvin = -273° Celsius), wird klar dass 100° Celsius nicht doppelt so heiss ist wie 50° Celsius, da 100° Celsius 373 Kelvin und 50° Celsius 323 Kelvin entspricht. Hingegen wäre 100 Kelvin doppelt so heiss wie 50 Kelvin. Solche Aussagen über *Verhältnisse* können erst auf dem höchsten Skalenniveau gemacht werden (vgl. Ingenkamp 1997, p. 31).

Ein weiteres Beispiel für eine Intervallskala wäre die Einteilung des Intelligenzquotienten (IQ) in einem sorgfältig standardisierten Intelligenztest (vgl. ebd., p. 32). Auch hier sind die Intervalle zwischen den einzelnen IQ-Werten gleich gross – d.h. der Abstand zwischen IQ 70 und IQ 140 ist gleich gross wie die Differenz zwischen IQ 60 und IQ 130 – nämlich jeweils 70 IQ-Einheiten. Wiederum wäre es aber falsch zu sagen, jemand mit IQ 140 sei doppelt so intelligent wie jemand mit IQ 70, weil der Nullpunkt auch hier willkürlich festgelegt wurde, respektive dieser gar nicht bekannt ist (vgl. ebd.).

Bei der Intervall- und auch der anschliessend behandelten Verhältnis- oder Proportionalskala spricht man auch von *metrischen Skalen*. Hier sind die üblichen algebraischen Operationen wie Addition oder Subtraktion der Werte sowie auch die Bildung von *arithmetischen Mittelwerten* erlaubt.

Auch die Grenze zwischen *qualitativen* und *quantitativen* Aussagen wird im Allgemeinen hier angesiedelt. Qualitative Daten erfüllen die Kriterien nominaler oder ordinaler Messungen, während quantitative Daten mindestens die Messung auf dem Intervallskalenniveau beanspruchen.

1.2.4 Verhältnis- oder Proportionalskala

Auf dem Niveau der *Verhältnis- oder Proportionalskala* sind nun auch Angaben über Verhältnisse wie «doppelt so gross», «halb so schwer» oder «viermal so breit» möglich. Solche Aussagen sind dann gestattet, wenn ein *natürlicher Nullpunkt* besteht – d.h. dieser nicht willkürlich an einem Ort der Skala festgesetzt wurde, sondern beim Skalenwert «0» das Merkmal auch tatsächlich Null ist. Solche Verhältnisse gibt es etwa bei Längen- oder bei Gewichtsmassen, oder auch die Kelvinskala bei der Temperaturmessung wäre ein solches Beispiel (vgl. Tabelle 1.1). Hier wären Aussagen wie 3 kg ist doppelt so schwer wie 1,5 kg zulässig.

Tabelle 1.1: Temperaturmessung nach Grad Celsius und Kelvin

	Absoluter Nullpunkt	***Gefrierpunkt von Wasser***		***Siedepunkt von Wasser***
Grad Celsius	-273° C	0° C	50° C	100° C
Kelvin	**0** K	273 K	323 K	373 K

Anmerkung: Die Zahlen sind gerundet. Ganz korrekt wäre der absolute Nullpunkt auf der Celsius-Skala bei –273.15° C.

Mit Daten auf Verhältnisskalenniveau können alle Arten von Operationen, also auch Quotienten oder Vielfaches, vorgenommen werden (vgl. Ingenkamp 1997, p. 32). Bei der Messung von pädagogisch oder psychologisch relevanten Merkmalen wird kaum je das Niveau einer Verhältnis- oder Proportionalskala erreicht. Ausnahmen wären etwa die Messung von Reaktionszeiten in einem Wahrnehmungsexperiment, was aber eher für Wahrnehmungspsychologinnen und -psychologen als für Lehrpersonen interessant sein dürfte, oder – auch nicht gerade spezifisch pädagogisch – Körpergrösse und Gewicht von untersuchten Personen. Meist werden wissenschaftliche pädagogisch-psychologische Daten also eher auf dem Intervallskalenniveau oder einem darunter liegenden Ska-

lenniveau dargestellt. Tabelle 1.2 fasst die vier Skalentypen, deren Aussagen und die erlaubten Verfahren noch einmal im Überblick zusammen.

Tabelle 1.2: Die vier Skalentypen, Aussagen, erlaubten Kennwerte und Beispiele in der Übersicht (angelehnt an: Fisseni 1997, p. 29; Ingenkamp 1997, p. 33)

Skala	***Aussagen***	***Erlaubte Kennwerte Erlaubte Verfahren***	***Beispiele***
Nominalskala	gleich / ungleich	Modalwert (häufigster Wert); Häufigkeiten	Augenfarbe, Geschlecht, Beruf der Mutter, Konfession
Ordinal- / Rangskala	grösser / kleiner mehr / weniger	Median (mittlerer Wert); Rangverteilungen	Schulnoten, Rangplätze bei Sportwettkämpfen
Intervallskala	Bestimmbarkeit der Skalenabstände; relativer Nullpunkt	Arithmetischer Mittelwert; Addition, Subtraktion	Temperaturskala nach Celsius, Intelligenzquotient, Kalenderdaten
Verhältnis- / Proportionalskala	Bestimmbarkeit von Proportionen; absoluter Nullpunkt	alle Arten von mathematischen Operationen	Längenmasse, Gewichtsmasse, Temperaturskala nach Kelvin

Als letzter Punkt soll angefügt werden, dass es zwar jeweils wünschenswert ist, auf einem möglichst hohen Skalenniveau zu messen. Wie Kleber (1992, p. 162) betont, hängt das Skalenniveau aber massgeblich vom *Messinstrument* ab. «Wenn das letzte nur ordinale Information liefert, so kann diese Information nie durch die Zuordnung von Zahlen, die Intervallskalenniveau vortäuschen, aufgewertet werden». Am häufigsten finden wir im *schulischen Alltag* diagnostische *Informationen auf dem Ordinalskalenniveau* – also Rangfolgen – vor.[24]

1.3 Grundzüge der klassischen Testtheorie

Im Folgenden werden die Grundzüge der *klassischen Testtheorie* dargestellt, welche in den fünfziger Jahren des letzten Jahrhunderts von Gulliksen (1950) systematisch zusammengestellt und weiterentwickelt wurde. Die Verfahrensweisen und Gütekriterien der klassischen Testtheorie sind bis heute von zentraler Bedeutung für die Entwicklung und Beurteilung psychologischer und sozialwissenschaftlicher Messverfahren (vgl. Wild & Krapp 2001, p. 523).[25]

Tent und Stelzl (1993, p. 41f.) weisen darauf hin, dass die «Testtheorie» nicht nur auf psychologische oder pädagogische, und seitens der Verfahren auch nicht nur auf «Tests», sondern auf alle Arten diagnostischer Informationsquellen, also z.B. auch auf Fragebogen, Beobachtungen oder Prüfungen, anwendbar sei. Fragen etwa nach der Genauigkeit der Messung oder dem Einfluss von Fehlern auf die Messung stellen sich nicht nur bei «pädagogisch-psychologischen Tests», sondern bei jeder Messung, «[...] so dass die klassische Testtheorie als ein allgemeiner begrifflicher Rahmen

[24] Um das Intervallskalenniveau zu erreichen bedarf es nach Kleber (vgl. 1992, p. 162) bestimmter Tests sowie umfangreicher und langfristiger theoretischer Bemühungen.

[25] Gewisse Grundannahmen der klassischen Testtheorie werden heute kritisiert und neuere Testtheorien (z.B. Probabilistische Testtheorie) theoretisch favorisiert. In der Praxis haben sich Tests, die auf neueren Testtheorien basieren, bisher jedoch noch nicht breit durchgesetzt. Für eine Kritik an der klassischen Testtheorie sei z.B. auf Amelang und Zielinski (1997, p. 61f.) verwiesen. Ausführlich werden die Grundzüge der klassischen Testtheorie in den gängigen Einführungen zur Pädagogischen oder Psychologischen Diagnostik (u.a. Tent & Stelzl 1993; Amelang & Zielinski 1997; Fisseni 1997) oder in Werken zum Testaufbau und zur Testanalyse (z.B. Lienert 1991) dargestellt. Die neueren Probabilistischen Modelle spielen u.a. bei der Diskussion um Kriteriumstests (lehr- und lernzielorientierte Tests) eine Rolle.

anzusehen ist, der es ermöglicht, die *Qualität von Messungen* zu beurteilen und die *Auswirkungen von Messfehlern* abzuschätzen» (ebd.; Hervorhebungen im Original).[26]

Im Wesentlichen dient die klassische Testtheorie zur Konstruktion und Bewertung von Erhebungsinstrumenten, welche *stabile Personenmerkmale* wie etwa Intelligenz oder Leistungsmotivation messen sollen (*Eigenschaftsdiagnostik*). Dabei sind in die Entwicklung der Testtheorie v.a. differentialpsychologisch orientierte Fragestellungen eingeflossen, d.h. vornehmlich die Frage nach Unterschieden *zwischen* Individuen zu einem Zeitpunkt (vgl. Wild & Krapp 2001, p. 528). Die zeitliche Stabilität eines Merkmals wird dabei vorausgesetzt.[27]

Von grundlegender Bedeutung für das Verständnis der klassischen Testtheorie ist die Unterscheidung zwischen beobachtetem Wert, wahrem Wert und Messfehler (vgl. Abbildung 1.3). Ein Grundgedanke ist, dass die Ergebnisse von Messungen nicht vollständig stabil sind, sondern *Zufalls-Zufallsschwankungen* unterliegen. Stellt man sich vor, dass ein Merkmal (z.B. Intelligenz) bei einer Person mehrere Male mit demselben Test gemessen wird, so würde man jedes Mal einen *beobachteten Testwert* – also das Ergebnis des IQ-Tests – erhalten, welcher mehr oder weniger schwanken würde. Der *wahre Wert* – also die «wahre Intelligenz» der Person – wäre derjenige Wert, welchen diese Person (hypothetisch) im Durchschnitt über eine *unendliche* Anzahl von wiederholten Messungen unter denselben Bedingungen erreichen würde (vgl. Tent & Stelzl 1993, p. 42f.).

Der wahre Wert bleibt schliesslich eine theoretische Grösse, die zwar statistisch geschätzt (z.B. anhand der *Vertrauensintervalle*), aber nie genau angegeben werden kann. Die Differenz zwischen dem beobachteten Testwert der Person und dem theoretisch definierten wahren Wert wird dabei als *Messfehler* bezeichnet. Man könnte auch sagen, dass der Messfehler als Grösse angesehen wird, die sich beim Vorgang des Messens über den «eigentlichen» oder wahren Wert legt und den beobachteten, durch die Testung erhaltenen Wert vom wahren Wert mehr oder weniger stark abweichen lässt. Der Messfehler überlagert also den wahren Wert (vgl. Amelang & Zielinski 1997, p. 34).

beobachteter Wert (einer Person X) = ***wahrer Wert*** (der Person X) + ***Messfehler***

Abbildung 1.3: Zusammenhang zwischen beobachtetem Wert, wahrem Wert und Messfehler in der klassischen Testtheorie

Messfehler gehen dabei einerseits auf alle unsystematischen und nicht kontrollierbaren Einflussgrössen, die auf ein Messergebnis einwirken und dieses verzerren können, zurück. Es können Einflüsse wie positive oder negative Stimmungen bei der untersuchten Person, Konzentrationsschwankungen, Ablenkung, nicht optimale Raumverhältnisse, Einflüsse, die vom Testleiter ausgehen etc. sein. Allgemein werden solche nicht kontrollierbaren Einflussgrössen auch als *Störvariabeln* bezeichnet. Andererseits gibt es einen Anteil an Messfehlern, der aufgrund der gewöhnlich nicht perfekten, also «unvollständigen» Zuverlässigkeit (*Reliabilität*) des Tests selbst entsteht. Auf die Reliabilität einer Messung werden wir etwas später in diesem Kapitel (vgl. Abschnitt 1.4.2) eingehender zu sprechen kommen.

---

[26] Die klassische Testtheorie besteht in ihrem Kern aus einer Sammlung von Aussagen, die nach dem Muster mathematischer Theorien aus einigen wenigen Axiomen abgeleitet und in den wesentlichen Punkten mathematisch-statistisch formuliert sind (vgl. Wild & Krapp 2001, p. 523). Auf die Darstellung von Formeln wird hier verzichtet.

[27] Veränderungen oder Prozesse (*Veränderungs- oder Prozessdiagnostik*) hingegen können mit dieser Theorie nicht befriedigend abgebildet werden.

Zurück zur Schätzung des wahren Wertes einer Person, was ja meist das Ziel eines Tests oder anderer diagnostischer Messverfahren ist.[28] Nehmen wir an, dass eine Person X bei einem Intelligenztest einen IQ-Wert von 100 erzielt hat, was genau dem Mittelwert entspricht, also einer «durchschnittlichen» Intelligenz. Dieser beobachtete Wert wird also durch einen Fehler überlagert und diese «Fehlerbehaftetheit» des Testergebnisses muss bei der Schätzung des wahren Wertes mitberücksichtigt werden. Der Bereich, in dem der wahre Wert mit einer bestimmten Irrtumswahrscheinlichkeit liegt, wird durch die Berechnung von *Vertrauensintervallen* (*Konfidenzintervalle*) angegeben. Nimmt man beispielsweise das 95%-Konfidenzintervall (es gibt auch ein 99%-Konfidenzintervall), so hiesse dies, dass sich mit 95-prozentiger Sicherheit aussagen liesse, in welchem *Bereich* sich der wahre Wert der Person befindet. Das Konfidenzintervall für den Intelligenztest HAWIK-III (vgl. Abschnitt 3.3) beispielsweise weist einen Vertrauensbereich bei 95% Wahrscheinlichkeit von +/- 6.20 IQ-Punkten auf. Ein Messergebnis von 120 Punkten sagt, dass der wahre IQ-Wert mit 95% Wahrscheinlichkeit zwischen 113,8 und 126,2 liegt.

Somit sollte ersichtlich geworden sein, dass Messergebnisse, auch wenn sie auf der Basis von sorgfältig entwickelten Instrumenten erhoben wurden, nicht als absolute Werte behandelt werden dürfen, da Messvorgänge immer mit einem gewissen Fehleranteil behaftet sind und vom Instrument, mit dem sie erhoben wurden, massgeblich abhängen. Auch sollten Mess- oder Testergebnisse unter Beizug des nötigen fachlichen Wissens (über das *Messinstrument* sowie auch über das zu messende *Phänomen*) interpretiert werden. Ein wichtiger Anhaltspunkt, um einen Test oder ein Verfahren sowie auch die Werte, die damit erhoben wurden, kritisch zu betrachten, sind die *Gütekriterien*. Diese werden im Folgenden dargestellt.

1.4 Gütekriterien

Aus der klassischen Testtheorie sind Kriterien abgeleitet worden, welche die Güte einer Messung, also die *Genauigkeit* und *Verlässlichkeit*, charakterisieren. Man spricht allgemein von *Gütekriterien*. Es sind dies die drei *Hauptgütekriterien Objektivität*, *Reliabilität* und *Validität*, sowie eine Reihe von sog. *Nebengütekriterien*, auf welche hier aber nicht eingegangen wird.[29]

1.4.1 Objektivität

Die *Objektivität* einer Messung ist dann gegeben, wenn die *Resultate möglichst unabhängig von der Person, welche die Messung vorgenommen hat, sind*. Anders ausgedrückt wäre eine Testung dann objektiv, wenn verschiedene Untersucher bei der Messung ein und desselben Merkmals zu demselben Ergebnis kommen würden. Dies ist wünschenswert, da mit einem Test Aussagen über die untersuchte Person und *nicht* über die Person des Untersuchenden (z.B. Testleiterin oder Testleiter oder auch Lehrperson) gemacht werden sollen.

Um subjektive Einflüsse bei der Beurteilung ein und desselben Tatbestandes zu verringern, müssen möglichst viele übereinstimmende Arbeitsschritte im Beurteilungsvorgang festgelegt, d.h. vereinheitlicht oder *standardisiert* werden. Dahinter steht die Annahme, dass durch *Standardisierung der*

---

[28] Wenn im Folgenden jeweils generell von «Tests» gesprochen wird, so können damit auch andere diagnostische Erhebungsverfahren wie Fragebogen, Interviews oder Prüfungen gemeint sein, weil sich mess- oder testtheoretische Annahmen über «Tests» (im engeren Sinne) auch auf andere Verfahren hin übertragen lassen.

[29] Nebengütekriterien sind etwa die *Ökonomie* eines Tests, also das Kriterium, ob ein Test ohne grossen zeitlichen und materiellen Aufwand, z.B. auch bei Gruppen (sog. Gruppentest) angewendet werden kann, oder die *Nützlichkeit* eines Tests, die dann gegeben ist, wenn ein hohes praktisches Bedürfnis zur Untersuchung des Verhaltens besteht und noch kein oder nur wenige Verfahren für diesen Zweck vorliegen (vgl. Ingenkamp 1997, p. 42f.).

Untersuchungsbedingungen die Situationseinflüsse konstant gehalten werden, d.h. die Messbedingungen für jede untersuchte Person oder bei derselben Person über verschiedene Zeitpunkte hinweg immer dieselben sind. Somit kann das registrierte Verhalten theoretisch allein als Funktion der Personmerkmale verstanden werden (vgl. Tent & Stelzl 1993, p. 30).

Diese Standardisierung kann auf drei unterschiedlichen Ebenen stattfinden: bei der Durchführung, der Auswertung und der Interpretation.

Durchführungsobjektivität bedeutet, dass das Testergebnis nicht davon abhängt, wer als Untersuchende oder Untersuchender einen Test mit dem Probanden durchführt. Um dies zu erreichen, werden alle möglichen Durchführungsschritte wie etwa die Anleitungen (Instruktionen*)* oder auch Hil-Hilfestellungen, die gegeben werden dürfen, vereinheitlicht und bei standardisierten Verfahren z.T. sogar wortwörtlich vorgegeben. Es wird festgelegt, welche Hilfsmittel benutzt werden dürfen und wie viel Zeit zur Verfügung steht. Werden solche Vorgaben nicht strikt eingehalten, sind die Resultate in Bezug auf die Normen im Prinzip nicht interpretierbar, da sie unter anderen als den vorgesehenen Bedingungen zustande gekommen sind.

Die *Auswertungsobjektivität* ist dann gegeben, wenn nach der Testdurchführung bei einem bereits vorliegenden Testprotokoll (z.B. Antworten der Probanden) das Testergebnis (z.B. IQ-Punktwert) nicht von der Person, welche den Test auswertet, abhängt. Bei standardisierten Verfahren ist sie etwa dann *nicht* gegeben, wenn einer Testleiterin oder einem Testleiter Auswertungsfehler wie etwa Zählfehler bei *geschlossenen Antwortformaten* (in Ratings oder Multiple-Choice-Aufgaben) unterlaufen. Auswertungsschablonen oder auch der Einsatz von Computerprogrammen können helfen, diese Art von Fehlern zu verhindern oder möglichst gering zu halten (vgl. Tent & Stelzl 1993, p. 44).

Etwas schwieriger ist es, die Auswertungsobjektivität zu garantieren, wenn Probanden Antworten selbst formulieren können oder der Spielraum möglicher Antworten gross ist (*offene Antwortformate*). In solchen Fällen müssen detaillierte Auswertungsregeln erarbeitet werden und die Auswertungsobjektivität muss (zumindest bei wissenschaftlichen Untersuchungen) empirisch überprüft werden.

Auch in der schulischen Praxis kommt der Auswertungsobjektivität eine zentrale Stellung zu. Wie Ingenkamp (1997, p. 35) betont, beziehen sich die meisten Untersuchungen über die Schwächen der traditionellen Leistungsbeurteilung auf die mangelnde Auswertungsobjektivität. «Wenn verschiedene Lehrer z.B. dieselbe Klassenarbeit unterschiedlich bewerten, dann werten sie nicht objektiv aus» (ebd.). Offene Antworten in Klassenarbeiten oder auch bei mündlichen Prüfungen erfordern deshalb eine inhaltliche Klassifikation (vgl. Wild & Krapp 2001, p. 525), d.h. es muss festgelegt werden, welche Kriterien nach welchen Massstäben bewertet werden.

Die Objektivität der Auswertung kann durch Korrelation[30] der Testwerte derselben Person, die durch mindestens zwei unterschiedliche Testleiterinnen oder Testleiter ausgewertet wurden, quantitativ bestimmt werden. Angaben zur Auswertungsobjektivität finden sich oft in den Testhandbüchern. Die Durchführungsobjektivität hingegen lässt sich quantitativ weniger gut überprüfen, da hierbei auch Einflüsse unzureichender Reliabilität miteinfliessen würden. Wenn die Bedingungen der Durchführung jedoch eingehalten werden, kann diese als hoch angesehen werden.

---

[30] Korrelation ist ein statistischer Fachbegriff für *Zusammenhang* oder «Assoziation» zweier Variabeln. Masse für die Stärke der Korrelation werden meist als Korrelationskoeffizienten (r) bezeichnet. Korrelationskoeffizienten können Werte zwischen minimal -1 und maximal +1 annehmen, wobei -1 einen perfekten negativen und +1 einen perfekten positiven Zusammenhang bezeichnet. Eine Korrelation von 0 würde bedeuten, dass gar kein Zusammenhang zwischen zwei Variabeln besteht (vgl. ILMES 2006).

Interpretationsobjektivität liegt dann vor, wenn verschiedene Untersuchende bei demselben vorliegenden Resultat zu denselben Schlussfolgerungen gelangen. Hilfreich können hierbei etwa Normtabellen sein, welche die Ausprägung des interessierenden Merkmals in Bezug auf die Normalpopulation verdeutlichen. Auch in der schulischen Praxis kann es sinnvoll sein, Testergebnisse nach vorliegenden Normtabellen in die Leistungsverteilung der entsprechenden Klassenstufe oder nach Lernzielkatalogen in das erreichte Lernniveau einzuordnen (vgl. Ingenkamp 1997, p. 36).

Meist hängen weitergehende Interpretationen und Schlussfolgerungen (z.B. Förderungsbedarf) jedoch von mehr als nur einem Testresultat ab. In der psychologischen Praxis (z.B. auf der Erziehungsberatung, EB) werden meistens mehrere Tests (sog. *Testbatterien*) zu unterschiedlichen Be-Bereichen durchgeführt und auch weitere diagnostisch relevante Informationen eingeholt (z.B. Gespräche mit Eltern, Lehrpersonen). Auch in der pädagogischen Praxis werden Entscheidungen wie z.B. Selektionsentscheidungen meist aufgrund mehrerer relevanter Informationsquellen und Datenbasen gefällt.

Abschliessend soll noch einmal auf die Relevanz des hier Gesagten für die pädagogische Praxis hingewiesen werden. Einschränkend soll angefügt werden, dass es in alltäglichen Unterrichtssituationen nicht möglich und wahrscheinlich auch nicht wünschenswert wäre, alle Eventualitäten vorherzusehen, alle Schritte genau zu planen und zu standardisieren und sich überdies peinlich genau an solche Vorgaben zu halten. Die (soziale) Unterrichtssituation ist dafür zu komplex und kann in dieser Komplexität auch nicht reduziert werden. Lehrpersonen werden sich immer wieder situationsbezogen und flexibel auf gewisse unvorhergesehene Situationen einstellen müssen. Dennoch sollte es gerade in der schulischen Praxis ein vorrangiges Ziel sein, *ein möglichst hohes Mass an Objektivität der Durchführung, der Auswertung oder auch der Interpretation* z.B. von Prüfungen (oder anderen Beurteilungsformen) *zu erreichen*, denn auch hier gilt, dass Verhalten (z.B. Messergebnisse wie Noten) nur mit früherem Verhalten (z.B. frühere Noten der Schülerinnen und Schüler bei der individuellen Bezugsnorm) oder dem Verhalten von anderen (Noten anderer Schülerinnen und Schüler bei der sozialen Bezugsnorm) verglichen werden kann, wenn es *unter weitgehend gleichen Bedingungen* erfasst, *ausgewertet* und aufgrund *derselben Massstäbe interpretiert* wird. Unzweifelhaft sind Ergebnisse eines Mathematiktests nicht vergleichbar, wenn nicht alle Schülerinnen und Schüler gleichviel Zeit zur Verfügung hatten oder nur einige einen Taschenrechner benützen durften. Im Endeffekt geht es hier um die sehr zentrale Frage der *Fairness* oder *Chancengleichheit aller Schülerinnen und Schüler*, welche anzustreben ein vorrangiges Ziel pädagogischer Bemühungen sein sollte.

1.4.2 Reliabilität

Die *Reliabilität* ist die *Zuverlässigkeit* oder *Genauigkeit* einer Messung. Die Überprüfung der Reliabilität einer Messung betrifft v.a. die Frage, wie sehr auf ein einmaliges Messergebnis vertraut werden kann resp. wie gross der geschätzte Messfehler ist, welcher den wahren Wert überlagert. So ist ein Metermass aus einer Platinlegierung reliabler als eines aus Eisen (weil sich dieses mit wechselnden Temperaturen ebenfalls verändert) oder ein Gummiband, das von Situation zu Situation mehr oder weniger gedehnt wird. Das Beispiel zeigt auch, dass eine noch so gute Skalenqualität (das Metermass liegt auf dem Niveau einer Proportionalskala) noch keine Garantie für eine qualitativ gute Messung ist.

Die Reliabilität einer Messung kann nicht aufgrund einer Einzelmessung bestimmt werden, sondern muss auf der Basis von repräsentativen Stichprobenerhebungen statistisch geschätzt werden. Ausgedrückt wird die Reliabilität mit einem Korrelationskoeffizienten «r» zwischen 0 und 1. Wäre die Reliabilität perfekt, d.h. gäbe es z.B. eine hundertprozentige Übereinstimmung zwischen zwei Testwiederholungen («retests») oder zwischen zwei Auswertern (*Interrater-Reliabilität*), dann wäre

die Reliabilität r = 1,0. Wäre überhaupt keine Übereinstimmung da, dann wäre sie r = 0. In der pädagogisch-psychologischen Praxis sind hundertprozentige Zuverlässigkeiten kaum anzutreffen.[31]

Es gibt (v.a. in der Testkonstruktion) Möglichkeiten, die Reliabilität eines Instruments zu schätzen und zu erfassen. Zur Ermittlung der *Wiederholungsreliabilität* (*Retest-Reliabilität*) lässt man, unter der Bedingung, dass mögliche Lerneffekte ausgeschlossen werden können, denselben Test von denselben Probanden zu verschiedenen Zeitpunkten bearbeiten und überprüft danach anhand von Korrelationen die Übereinstimmung zwischen den Ergebnissen der beiden Messzeitpunkte.

Für gewisse Leistungsmessungen eher angebracht ist die Ermittlung der *Paralleltestreliabilität*. Voraussetzung ist, dass zwei parallele Testversionen (z.B. Version A und B eines Tests) vorliegen, die einander in ihren Aufgabenstellungen möglichst ähnlich sind. Nach Durchführung beider Versionen, die auch sehr kurz aufeinander folgen kann, wird wiederum die Übereinstimmung anhand eines Reliabilitätskoeffizienten ausgedrückt.

Eine dritte Möglichkeit, und dies ist ev. die geläufigste Methode, ist die Ermittlung der *Split-half-Reliabilität* (*Halbierungsmethode*) oder die Ermittlung der *inneren Konsistenz* (*Konsistenzanalyse*). Hier wird ein Test in zwei Hälften aufgeteilt, z.B. nach geraden und ungeraden Aufgabennummern (*Odd-Even-Methode*). Danach wird wiederum die Übereinstimmung der ersten mit der zweiten Hälfte des Tests anhand der Berechnung des statistischen Zusammenhangs überprüft. Bei der Konsistenzanalyse wird der Test statistisch in so viele Teile zerlegt, wie er Aufgaben (*Items*) enthält. Die Übereinstimmung der Items (auf der Basis der Homogenität der Antworten) wird dabei häufig durch die Angabe von *Cronbach's alpha* angegeben. Die Höhe dieser Reliabilität (z.B. pro Skala) wird meist in den Handbüchern angegeben und bietet dem Testbenutzer eine gute Möglichkeit, die Messgenauigkeit des Verfahrens abzuschätzen. Dieses Verfahren kann auch bei einer einmaligen Erhebung in einer Stichprobe durchgeführt werden.

Für Lehrpersonen ist die Kenntnis des Reliabilitätskoeffizienten eines Tests ev. von geringerem Wert als das *Vertrauensintervall*, innerhalb welchem sich der «wahre Wert» mit einer bestimmten Wahrscheinlichkeit befindet. Arbeiten von Ingenkamp lassen die Vermutung zu, dass der Messfehler und damit das Vertrauensintervall für das Zensurensystem bei plus/minus einer halben Notenstufe (für schweizerische Verhältnisse) anzusetzen ist: Hätte man also für eine bestimmte Schülerin bei einer Prüfung die Note 5 ermittelt, müsste man annehmen, dass die «wahre Note» dieser Schülerin tatsächlich irgendwo im Bereich zwischen 4.5 und 5.5 liegt. Diese Schwankung kommt alleine durch die mangelnde Zuverlässigkeit des Beurteilungsverfahrens zustande.

1.4.3 Validität

Das wichtigste Gütekriterium für ein Untersuchungs- oder Messverfahren ist dessen *Validität* oder *Gültigkeit*. Sie wird im Anschluss an Lienert (1969, p. 12) als «der Grad der Genauigkeit, mit dem ein Test dasjenige Persönlichkeitsmerkmal, das er messen soll oder zu messen vorgibt, tatsächlich misst» definiert. Die Gültigkeit oder Validität eines Verfahrens sagt also aus, *ob das Verfahren auch dasjenige Merkmal erfasst, das es messen soll, und nicht irgend etwas anderes*.

---

[31] Eine Reliabilität von r = 0.7 ist in der Praxis für die Beurteilung individueller Differenzen ausreichend, r = 0.8 ist akzeptabel (vgl. Lienert & Raatz 1994, p. 269). Eine Reliabilität von r = 0.8 besagt, dass 80% der beobachteten Testwertevarianz auf wahre Unterschiede zwischen den Personen zurückzuführen sind, und zu 20% auf der Fehlervarianz beruhen. Häufig wird das Mass R^2 (Korrelationskoeffizient im Quadrat) als Mass für die «aufgeklärte Varianz» angegeben. R^2 gibt an, welcher Anteil der Varianz der abhängigen Variable (= Kriterium) durch die unabhängige(n) Variable(n) (= Prädiktoren) erklärt wird. Bei einem r = 0.8 wäre dieses also $R^2 = 0.64$ und würde bedeuten, dass 64% der Varianz der abhängigen Variable durch die unabhängige(n) Variable(n) erklärt oder vorausgesagt werden kann (vgl. ILMES 2006).

Wie eingangs bereits erörtert, ist die Bestimmung der Validität diagnostischer Verfahren ein besonderes Problem, da sich Messungen in der Regel auf ein nicht direkt beobachtbares *Konstrukt* (z.B. Intelligenz, Prüfungsangst) beziehen und deshalb eine unmittelbare Überprüfung der Validität nicht möglich ist (vgl. Wild & Krapp 2001, p. 525f.). Am einfachsten scheint die Frage nach der Validität dann beantwortet zu sein, wenn die Testaufgaben selbst eine Stichprobe aus dem Verhaltensbereich sind, über den eine Aussage getroffen werden soll. Dies ist etwa der Fall, wenn sich jemand zur Beurteilung seiner Fahrtauglichkeit während der praktischen Fahrprüfung eine Stunde lang im Verkehr beweisen muss oder wenn etwa zur Überprüfung der Rechtschreibekenntnisse ein Diktat durchgeführt wird. In solchen Fällen spricht man von *inhaltlicher* oder auch *logischer Validität*.

Im pädagogischen Bereich stellt sich die Frage nach der *Inhaltsvalidität* (Kontentvalidität) etwa bei der Konstruktion von Schulleistungstests oder schriftlichen Arbeiten. Hierbei gilt es festzustellen, ob der Pool von Aufgaben, der z.B. zur Erfassung eines Wissensgebietes vorliegt, dieses auch hinreichend genau repräsentiert (vgl. Wild & Krapp 2001, p. 526). Eine Lehrperson konstruiert für eine schriftliche Prüfung Fragen, aus deren richtiger Beantwortung sie auf den spezifischen Lernerfolg schliessen kann. Dabei wird häufig übersehen, dass eine angemessene Beschreibung des inhaltlichen Gültigkeitsbereichs eines Verfahrens schwieriger ist, als es auf den ersten Blick zu sein scheint resp. unterschiedliche Untersuchende oder Lehrpersonen sich in der Auswahl oder der Konstruktion von Items, welche für ein Gebiet repräsentativ sein sollen, durchaus unterscheiden können.

Ingenkamp (1997, p. 40f.) weist in diesem Zusammenhang auf die *curriculare Gültigkeit* (als Spezialfall der Inhaltsgültigkeit) hin, welche durch die Vorgabe *operationalisierter Lernziele* erreicht werden kann. Gemeint ist damit, dass die in Lehrplänen beschriebenen Lernziele und thematischen Anforderungen nicht nur abstrakt, sondern inhaltlich möglichst klar beschrieben werden sollten.[32] Auch ist es sinnvoll, wenn in den Curricula festgehalten wird, unter welchen Bedingungen ein angestrebtes Endverhalten gezeigt werden soll sowie auch, über welche notwendige Lehrtätigkeit dieses Endverhalten erreicht werden kann. Zudem ist es zur Verringerung von Interpretationsdifferenzen sinnvoll, wenn festgelegt wird, welcher Massstab zur Messung des angestrebten Endverhaltens angewendet werden soll. «Die so *operationalisierten Lernziele* [...] können direkt in Prüfungsaufgaben umgesetzt werden» (ebd.; Hervorhebungen im Original).[33]

Bei der Ermittlung der *Kriteriumsvalidität* geht es um die Frage, ob die Resultate eines diagnostischen Verfahrens auch mit anderen, objektiven Aussenkriterien korrespondieren resp. nicht korrespondieren (vgl. Wild & Krapp 2001, p. 526). So kann ein Test zur sozialen Intelligenz zusätzlich etwa mit einer Verhaltensbeobachtung, Befragung der Schulkameraden oder der Eltern, Lehrerinnen- und Lehrerurteilen oder mit anderen Tests, die ebenfalls soziale Intelligenz messen, verglichen werden. Eine weitere Möglichkeit die Kriteriumsvalidität zu bestimmen ist die *Vorhersagevalidität* (*prädiktive* oder *prognostische Validität*). Diese ist von grossem Interesse, speziell im Hinblick auf Selektionsentscheidungen, da überprüft wird, inwiefern ein prognostiziertes Verhalten später tatsächlich auch eintritt. Dabei wird eine Messung zu einem ersten Zeitpunkt gemacht, bei dem ein Untersuchungsbefund festgehalten und allenfalls eine Entscheidung getroffen wird (z.B. Schulreife oder Zulassung zum Medizinstudium). Später wird der Zusammenhang dieses Befundes mit ausge-

---

[32] Eine neuere Entwicklung im Rahmen des Prozesses zur Harmonisierung der obligatorischen Schule in der Schweiz ist die Festlegung der verbindlichen Kompetenzen der Schülerinnen und Schüler durch sogenannte *Standards* (vgl. EDK 2003a).

[33] Eine gute Darstellung von Beispielen operationalisierter Lernziele findet sich in Gage und Berliner (1996, p. 32ff.). Es wird etwa für das Fach Geschichte das Lernziel «Ohne Hilfsmittel beschreibt der Schüler jeweils fünf Aspekte, welche die Gemeinsamkeit kapitalistischer und sozialistischer Systeme charakterisieren, und fünf, die diese Systeme unterscheiden» beispielhaft genannt, oder für den Bereich Algebra «Der Schüler kann in 20 Minuten ohne weitere Hilfsmittel 8 von 10 Gleichungen des Typs $AX^2+BX+C=0$ lösen» (ebd., p. 35).

suchten Kriterien (z.B. Schulerfolg, Studien- und Berufserfolg) errechnet. Die Vorhersagevalidität eines Verfahrens ist dann gegeben, wenn es bei der ersten Testung diejenigen Personen z.B. als «schulreif» erkennt, die dann auch tatsächlich in der Schule erfolgreich sind resp. das Kriterium «Schulerfolg» erfüllen.

Abschliessend zu den Ausführungen über die Gütekriterien soll auf deren gegenseitigen *engen Zusammenhang* hingewiesen werden (vgl. Abbildung 1.4). Das wichtigste Kriterium, die Gültigkeit oder Validität einer Messung, ist nur gegeben, wenn die Messung zuverlässig und genau, d.h. reliabel ist. Dies wiederum ist nur möglich, wenn die Ergebnisse unter objektiven Bedingungen zustande gekommen sind, oder anders gesagt: *«Ohne ein Mindestmass an Objektivität ist keine Reliabilität denkbar, ohne befriedigende Reliabilität keine hinlängliche Validität»* (Amelang & Zielinski 1997, p. 183 – eigene Hervorhebungen).

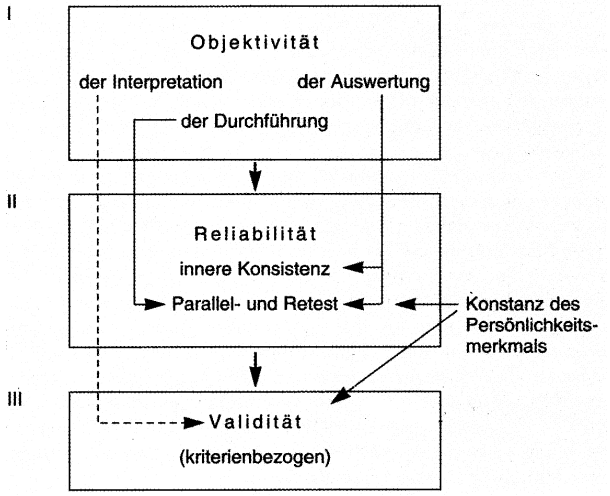

Abbildung 1.4: Wechselbeziehungen zwischen den Hauptgütekriterien Objektivität, Reliabilität und Validität (leicht modifiziert aus: Amelang & Zielinski 1997, p. 184)

1.5 Normen

In der pädagogischen Diagnostik werden, wie wir gesehen haben, Informationen zur Beurteilung eines Merkmals bei einer Person gesammelt. Damit solche Informationen ausgewertet und interpretiert werden können, ist es sinnvoll, *Normen* als Vergleichsbasis zur Verfügung zu haben. Wenn wir beispielsweise wissen, dass eine Person in einem bestimmten Test 27 Punkte erreicht hat oder ein Schüler in einem Diktat 12 Fehler geschrieben hat, so wissen wir noch nicht viel darüber, wie dieses Resultat zu interpretieren ist, weil die *Bezugspunkte* oder die *Massstäbe* fehlen. Wir müssen zusätzliche Informationen darüber haben, wo z.B. der Mittelwert des Tests lag oder wie andere Schülerinnen und Schüler mit ähnlichen Voraussetzungen in diesem Diktat abgeschnitten haben, bzw. wie viele Fehler erlaubt sind, um das Lernziel zu erreichen (vgl. Ingenkamp 1997, p. 44).

Dabei gilt es vorerst zu unterscheiden, mit welcher *Art von Normen* ein Ergebnis verglichen wird. Bei der *sozialen Bezugsnorm* wird das Ergebnis der Messung einer Merkmalsausprägung bei einer Person mit dem Ergebnis anderer Personen verglichen. Hat eine Schülerin in einem Mathematiktest 28 Punkte erzielt und liegt der Mittelwert der Leistungen aller Geprüften bei 20 Punkten, so hat diese Schülerin in diesem Test überdurchschnittlich gut abgeschnitten. Die gleiche Punktzahl erhält eine ganz andere Bedeutung, wenn der Mittelwert der Leistungen z.B. 35 Punkte betragen hätte.

Bei der *individuellen* Bezugsnorm spielt das Abschneiden oder die Leistung anderer Personen keine Rolle. Hier wird nur die individuelle Leistungsentwicklung eines Schülers zur Grundlage der Messwertinterpretation gemacht. Hat ein Schüler in einem vorausgegangenen, vergleichbaren Test

nur 15 Punkte erzielt, dann sind seine jetzt erreichten 28 Punkte als Fortschritt oder als «gut» zu bewerten. Nicht interpretieren können wir allerdings Resultate, die aus nicht vergleichbaren Tests stammen. Wenn wir wissen, dass ein Schüler beim ersten Diktat 12 Fehler und bei einem zweiten 18 Fehler gemacht hat, so können wir nicht sagen, ob dies eine individuelle Verbesserung oder Verschlechterung der Leistung ist, solange wir nichts über die *Schwierigkeit* des Diktats wissen. Für die soziale und die individuelle Bezugsrichtung sind deshalb Vergleichsnormen die besten Interpretationshilfen (vgl. Ingenkamp 1997, p. 44).

Eine dritte Möglichkeit ist die *kriteriale* oder *kriteriumsbezogene Bezugsnorm*. Hier wird die Leistung in Bezug auf ein vorher festgelegtes Kriterium, z.B. gesetzte Lernziele, bewertet. Im obigen Beispiel könnte es sein, dass festgelegt wird, dass von den maximal 60 erreichbaren Punkten mindestens 70% hätten erreicht werden müssen. Somit ist der Kandidat mit 28 Punkten leider durch die Prüfung gefallen.

Es wird deutlich, dass gleiche Resultate als Leistungen ganz anders in Erscheinung treten oder verschiedene Folgen haben können, je nach dem, welche Bezugsnorm angewendet wird (vgl. Rheinberg 1998, p. 39). Bemerkenswert ist dabei, dass sich (nicht-professionelle) Beurteilerinnen und Beurteiler meist darüber im Unklaren sind, welche Bezugsnorm sie vermeintlich selbstverständlich oder «natürlicherweise» anlegen. Empirisch konnte gezeigt werden, dass sich Personen in der Bevorzugung einer Bezugsnorm unterscheiden. Das Phänomen der *Bezugsnormorientierung* wurde speziell auch bei Lehrpersonen untersucht (z.B. Rheinberg 1980). In einer Reihe von Untersuchungen konnte gezeigt werden, dass Lehrpersonen mit einer starken sozialen Bezugsnormorientierung ungünstige Motivationseffekte (z.B. auf das schulische Selbst- und Fähigkeitskonzept) insbesondere bei lernschwächeren Schülerinnen und Schülern erzeugen können (vgl. Rheinberg et al. 2001, p. 314).

Wie bereits angesprochen wurde, ist das Heranziehen quantitativer Normen besonders für die Interpretation von Leistungen in Bezug auf soziale oder individuelle Vergleiche sinnvoll. Unter dem Blickwinkel sozialer Bezugsnormen wäre eine Leistung dann sehr gut, wenn nur sehr wenige andere Personen derselben Bezugsgruppe (z.B. derselben Alters- oder Klassenstufe) bessere oder gleich gute Resultate erzielt haben. Häufig wird die relative Position eines Merkmalsträgers (in Bezug auf die Gesamtpopulation) ausgedrückt durch *Prozenträngе*. Hätte beispielsweise ein Schüler bei einem Rechentest für die 4. Klasse einen Prozentrang von 75, so würde dies bedeuten, dass er gleich oder besser (≥) abgeschnitten hat als 75% aller Schülerinnen und Schüler von 4. Klassen (vgl. Ingenkamp 1997, p. 45).

In standardisierten Tests liegen meist *Normtabellen* vor, anhand derer die Rohwerte in Normwerte umgerechnet werden können. Zur Ermittlung solcher Normtabellen, die jeweils zuhinterst in den Testhandbüchern angegeben sind, müssen vorerst grosse und *repräsentative Eichstichproben* untersucht werden. Repräsentativ heisst, dass die Eichstichprobe ein verkleinertes Abbild der Grundgesamtheit[34] sein sollte, für welche die Normen gelten sollen. Anhand der Verteilung der Rohwerte bei der Eichstichprobe werden Normen, meist separat für jede Altersstufe, je nach Test (z.B. bei standardisierten Schulleistungstests auch nach Schulstufe und Regionen) oder separat für die Geschlechter, errechnet. Angaben über die Zusammensetzung oder auch die Grösse der Eichstichprobe sollten in den Handbüchern gemacht werden. Vor allem bei sehr speziellen Personengruppen (z.B. bei klinisch auffälligen Personen oder in Leistungstests bei sehr hohen/niedrigen Werten) können

[34] In der empirischen Forschung bezeichnet die Grundgesamtheit (auch Grundpopulation) die Menge aller potentiellen Untersuchungsobjekte für eine bestimmte Fragestellung. Aus pragmatischen Erwägungen wird normalerweise nicht die Grundgesamtheit, sondern eine Stichprobe untersucht, die für die Grundgesamtheit repräsentativ ist.

die Testnormen infolge der sehr kleinen Vergleichsstichprobe, aufgrund derer die Normen ermittelt wurden, ev. nicht uneingeschränkt interpretiert werden. Da die Datenerhebung für die Testnormierung sehr aufwändig ist und deshalb viele Tests in diesem Punkt Mängel aufweisen, hat man bei der Beurteilung der Qualität eines Tests auf die *Qualität der Normen* besonders zu achten.

Wenn sich die an einer repräsentativen Stichprobe ermittelten Rohwerte *normal* verteilen, können sie in standardisierte *z-Werte* (Mittelwert 0, Standardabweichung 1) umgerechnet werden und ihnen die entsprechenden Prozentränge der *Standard-Normalverteilung* zugeordnet werden (vgl. ebd., p. 57f.). Die Standard-Normalverteilung ist eine symmetrische Kurve (Gauss'sche Glockenkurve), bei der oberhalb und unterhalb des Mittelwertes genau 50% der Verteilung der Werte liegen. Abweichungen vom *Mittelwert* (M) werden mit der *Standardabweichung* (S) gekennzeichnet. Innerhalb einer Standardabweichung (über und unter dem Mittelwert) liegen ungefähr 68% der Messwerte aller Personen. Abbildung 1.5 kann helfen, dies zu verdeutlichen.

Bei der Normierung von *Abweichungs-Intelligenzquotienten (IQ)* werden meist ein Mittelwert von 100 und eine Standardabweichung von 15 angesetzt.[35] Dies bedeutet, dass die IQ-Normierung so angelegt ist, dass 68% aller Personen der Grundgesamtheit einen gemessenen IQ zwischen 85 und 115 haben und nur gerade je 16% darunter resp. darüber liegen. Weiter kann ein Resultat wie IQ 115 (ohne die Vertrauensintervalle vorerst mit zu berücksichtigen) in den Prozentrang 84 umgerechnet werden, was bedeutet, dass diese Person gleich gut oder besser wäre als 84% ihrer Vergleichsgruppe. Meist ist es einfacher und handlicher, z- oder IQ-Werte in die entsprechenden Prozentrangwerte umzurechnen, um sie zu interpretieren.

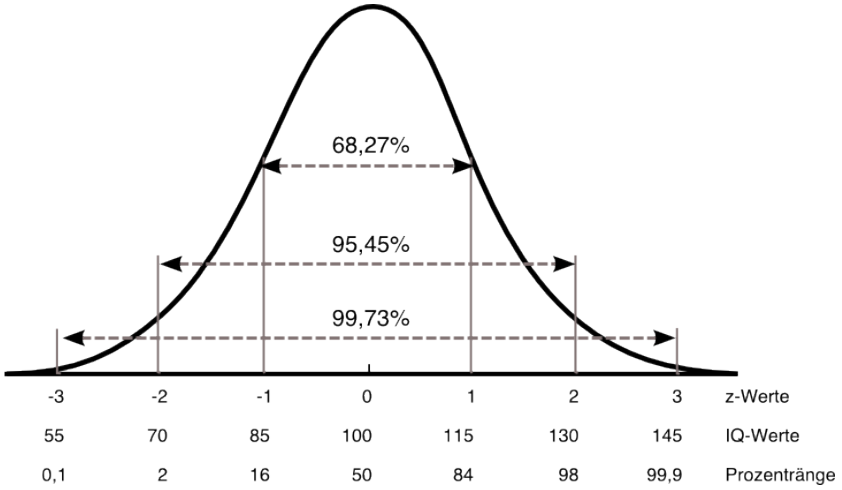

Abbildung 1.5: Normal verteilte Testwerte und verschiedene gebräuchliche Normskalen (in Anlehnung an: Wild & Krapp 2001, p. 531)

In gewisser Hinsicht zwar problematisch aber dennoch gebräuchlich sind Einteilungen und *Klassifikationen von Begabung* nach quantitativen Gesichtspunkten (in der Regel nach Intelligenz). Von *Hochbegabung* wird bei dieser – wie schon gesagt rein quantitativen Definition – dann gesprochen, wenn jemand einen IQ erreicht, der 2 Standardabweichungen (vgl. oben in der Grafik die z-Werte) über dem Mittelwert liegt. Dies ergäbe folglich einen IQ ≥ 130. Diese Person wäre also intelligenter als 98% ihrer Vergleichsgruppe. Etwa in derselben Grössenordnung liegt übrigens auch der mittlere Prozentsatz, den Lehrpersonen bei der Frage nach der Vorkommenshäufigkeit hochbegabter Schülerinnen und Schüler in ihren Klassen angeben (vgl. Rost 1998, p. 177). Von «Höchstbegabung» oder «Genie» wird ab einem IQ von 160 gesprochen. Wenn Sie wiederum oben schauen, sehen Sie, dass es in dieser Abbildung für einen so hohen IQ gar keine Angaben mehr gibt – diese Person –

[35] Intelligenz und Intelligenztests werden in Kapitel 3 ausführlicher diskutiert.

vielleicht Albert Einstein oder Marie Curie – wäre aber intelligenter als 99.9% ihrer Vergleichspopulation.

Was das Thema Hochbegabung nebst seiner rein quantitativen Bestimmung sonst noch für Implikationen haben könnte, wird inhaltlich ausführlicher und anhand von Fallbeispielen im Vertiefungsteil zur Begabungsdiagnostik diskutiert (vgl. Kapitel 3). Dort wollen wir es auch nicht versäumen, gewisse Kritikpunkte an der Intelligenzdiagnostik zu formulieren – eine Perspektive, welche in diesem methodischen Grundlagenkapitel bisher noch wenig zur Sprache gekommen ist. Vorerst wenden wir uns aber einem weiteren für den pädagogischen Alltag relevanten Thema zu: der schulischen Leistungsdiagnostik und Notengebung.

2 Schulische Leistungsdiagnostik und Notengebung

2.1 Fallbeispiel

Montagabend, Frau Lutz sitzt in ihrem Schulzimmer und bereitet ihren Unterricht vor. Das Thema wird das Berner Oberland sein. Die Lehrperson fragt sich, auf welche Art und Weise sie den Stoff präsentieren und wie sie das Gelernte im Anschluss prüfen will. Soll sie mit Vorträgen und Arbeitsblättern arbeiten und am Schluss eine Prüfung machen? Soll sie eine Werkstatt aufbauen und die Schülerinnen und Schüler ein Portfolio erstellen lassen? Wie kann die Lehrperson am besten herausfinden, ob die Klasse das Wichtigste gelernt hat? Welcher Beurteilungsmassstab wird ihnen am ehesten gerecht? Auf welche Art und Weise können die Schülerinnen und Schüler *für ihr Lernen* und die Lehrpersonen *für ihren Unterricht* am meisten von der Leistungsbeurteilung profitieren?

Wie dieses Fallbeispiel und das folgende Zitat von Ingenkamp zeigen, hat schulische Leistungsdiagnostik zwei zentrale Ziele: Lernvorgänge messen und sie für das pädagogische Handeln nutzbar machen:

> «Schulleistungstests sind Verfahren der Pädagogischen Diagnostik, mit deren Hilfe Ergebnisse geplanter und an Curricula orientierter Lernvorgänge möglichst objektiv, zuverlässig und gültig gemessen und durch Lehrende (z.T. auch durch Lernende) oder Beratende ausgewertet, interpretiert und für pädagogisches Handeln nutzbar gemacht werden können» (Ingenkamp 1997, p. 117).

In den folgenden Abschnitten wird den damit verbundenen Fragen nachgegangen. Im ersten Teil wird der schulische Leistungsbegriff bestimmt und eingegrenzt (Abschnitt 2.2). Im Abschnitt 2.3 wird aufgezeigt, was unter Diagnosekompetenz von Lehrpersonen verstanden wird, und es werden einige empirische Ergebnisse zur Diagnosekompetenz dargestellt. In den drei folgenden Teilen werden verschiedene diagnostische Verfahren vorgestellt: nicht-standardisierte (informelle) (Abschnitt 2.4) und standardisierte Schulleistungstests (Abschnitt 2.6). Dazwischen wird der Notengebung als zentrales Instrument der Leistugsbeurteilung ein eigener Abschnitt (2.5) gewidmet. Darin werden einerseits die wichtigsten Kritikpunkte an der Notengebung und Verbesserungsmöglichkeiten aufgeführt, andererseits wird erläutert, weshalb die Notengebung trotz Kritik weiterhin ein wichtiges Instrument der Leistungsrückmeldung bleibt. In Abschnitt 2.7 werden am Beispiel von «Basis-Check» und «Multi-Check» Aspekte der Leistungsbeurteilung durch ausserschulische Instanzen thematisiert. Im abschliessenden Abschnitt 2.8 wird ein Fazit gezogen und es wird der Versuch unternommen, einige Grenzen der schulischen Leistungsdiagnostik zu erläutern.

2.2 Schulische Leistungsbeurteilung

Die schulische Leistungsbeurteilung basiert auf der Beurteilung von Menschen durch Menschen. Eine Beurteilung muss die Individualität einer Person berücksichtigen. Gleichzeitig wollen Menschen in ihrer Sozialität wissen, wo sie im Vergleich zu anderen steht. Individualität und Sozialität stehen sich gegenüber und konfrontieren die Schule mit widersprüchlichen Herausforderungen:

- Beurteilung dient nicht nur dem Lernprozess des Individuums, sondern verfolgt auch andere Funktionen. Sie ist in dem Sinne nicht nur persönlich, sondern auch öffentlich. Dies zeigt sich dann, wenn nicht nur gefördert, sondern auch selektioniert werden muss.
- Eine schulische Beurteilung sagt nicht nur etwas über die Schülerinnen und Schüler aus, sondern auch etwas über ihr soziales Umfeld sowie die Lehrpersonen, die den Unterricht durchführen. Dennoch kommen die Konsequenzen, die aus schulischen Beurteilungen resultieren, hauptsächlich den Beurteilten zugute beziehungsweise werden von ihnen getragen.

2.2.1 Der schulische Leistungsbegriff

Das Wort *leisten* hatte in früheren Jahrhunderten die Bedeutung von *nachspüren, lehren, lernen* und *wissen* (vgl. Ingenkamp 1997, p. 95ff.). Um 1930 wurde der Begriff *Leistung* in pädagogische Lexika aufgenommen, wo er durch die steigenden schulischen Anforderungen und die Entwicklung der Schulen als gesellschaftliche Institution in vierfacher Bedeutung gebraucht wurde (vgl. ebd. p. 96):

- Leistung als schulische Förderung des Schülers resp. der Schülerin;
- Leistung als Tätigkeit des Schülers resp. der Schülerin;
- Leistung als Ergebnis der Tätigkeit des Einzelnen innerhalb verschiedener Leistungsbereiche;
- Leistung als besonderer Beitrag der Schule für Gesellschaft, Staat, Wirtschaft und Wissenschaft.

Um 1970 wurde der Begriff *Leistung* im pädagogischen Kontext sehr kontrovers diskutiert. Leistungsschule war für viele zu einem Abbild der entarteten Leistungsgesellschaft geworden. Die Bezeichnung «Leistung» wurde in Beziehung gesetzt zur kapitalistischen Dienstleistung und Warenproduktion (vgl. Ingenkamp 1997, p. 95ff.). Lernen in der Schule ist aber etwas anderes als Dienstleistungen oder Warenproduktion in der Wirtschaft. Was mit dem kapitalistischen Prinzip als gemeinsam angesehen werden kann, ist die Abhängigkeit des Ansehens und der Entlöhnung eines Menschen von seinen erbrachten Leistungen. Sie gilt als die Basis der demokratischen Leistungsgesellschaft. Pädagogisch gesehen ist die Leistung eines Schülers oder einer Schülerin *das Ergebnis seiner/ihrer Bemühungen, die Anforderungen des Unterrichts zu erfüllen*. Diskussionen über die Anforderungen des Unterrichts wurden verstärkt aufgenommen, als Schulleistungen einen bestimmenden Faktor für den sozialen Aufstieg wurden.

In der Beurteilung der Lernerfolge muss zwischen konvergenten, divergenten sowie emotionalen und sozialen Leistungen unterschieden werden (vgl. ebd. p. 97). Als *konvergente* Leistungen werden jene Leistungen bezeichnet, die sich in richtigen oder eindeutig besten Ergebnissen ausdrücken lassen; dies reicht vom Faktenwissen bis zur begründeten Wertung (Berichte, Beschreibungen usw.). Anders ist es bei *divergenten* Leistungen (Stimmungsbilder, kreatives künstlerisches Schaffen usw.), die zu vielen gleichwertigen Ergebnissen führen können, und sich meist der Messung, auch der kriteriumsorientierten, entziehen. Wenn man bestimmte divergente Leistungen unverzichtbar zum Nachweis des Lernerfolgs benötigt, dann muss überlegt werden, wie eine sachgerechte Bewertung am besten möglich ist (bspw. Beurteilung mit Portfolio, vgl. Abschnitt 2.4.4). Auch die Erreichung *sozialer und emotionaler* Lernziele bedarf einer spezifischen Evaluation. Beobachtungsmethoden,[36] Interviews, Schätzskalen und in einigen Fällen auch Tests können diesbezüglich angemessene Verfahren sein.

Auf die Frage, wer die Lernerfolge misst und bewertet, gibt es verschiedene Antworten: So gibt es für die Schulabschlussqualifikationen (in Deutschland und der Schweiz) interne Examen. Grossbritannien praktiziert ein überwiegend externes Prüfungssystem, Holland eine Mischform (vgl. ebd., p. 98).

Somit stehen sich unterschiedliche Institutionen und Individuen mit unterschiedlichen Zielsetzungen, verschiedenen Leistungstypen (divergent, konvergent, emotional-sozial) und Lernerfolgsmessungen (intern, extern, Mischform) gegenüber, die innerhalb der schulischen Leistungsdiagnostik berücksichtigt werden müssen. Fest steht, dass keine Einheitsmethode schulischer Leistungsbeurteilung allen Zielsetzungen gerecht wird. Es braucht unterschiedliche Verfahren, je nach Zielsetzung und Inhaltsbereich. Kritisiert wird in diesem Zusammenhang etwa das einseitige Übergewicht des kognitiven Bereichs bei der Beurteilungspraxis (vgl. Ingenkamp 1997,

[36] vgl. Kapitel 4: Unterrichtsbeobachtung.

p. 101) und innerhalb dieses kognitiven Bereichs die weitgehende Orientierung an Erinnerungsaufgaben. Diese Problematik wird im folgenden Abschnitt erörtert.

2.2.2 Leistung zwischen Erinnern und Denken

Wissensaufgaben basieren eher auf *Erinnern* bereits gemachter Erfahrungen resp. auf dem deklarativen, faktenspezifischen Gedächtnis (vgl. Haberlandt 1999, p. 69ff.). Attribute wie *Denken* und *Verstehen*, die mit dem intuitiven und unbewusst laufenden prozeduralen Gedächtnis in Verbindung stehen (vgl. ebd., p. 103), werden bei Verständnis- und Anwendungsfragen in den Vordergrund gerückt. Bei der Analyse von 3827 Prüfungsarbeiten aus kaufmännischen Berufsschulen stellten Forschende fest, dass 93-97% aller gestellten Aufgaben Wissensaufgaben waren und sich nur 3-7% auf Verständnis und Anwendung bezogen (vgl. Ingenkamp 1997, p. 121). Erziehen wir unsere Schülerinnen und Schüler in der Schule zu Erinnerungsmaschinen ohne Verständnis?

Im Kontext der Schule wird eine stärkere Betonung von Verständnis- und Anwendungsfragen gegenüber blosser Erinnerungsaufgaben gefordert. Ein Paradigma, mit dem sich die Schule aktiver auseinanderzusetzen hat. Autoren wie Adams (1989) und Messner (1978) legen grossen Wert auf die Unterscheidung zwischen *abstraktem Wissen* und *abstrahiertem Wissen*. Abstraktes Wissen bleibt sehr oft unbrauchbar, träge. Es lässt sich im konstruktiven Prozess des Wissenserwerbs oder auch des Problemlösens nicht mit anderen begrifflichen Einheiten verbinden. Im Gegensatz dazu steht das abstrahierte Wissen, das sich mit andern begrifflichen Einheiten verbinden und somit auf neue Lernsituationen transferieren lässt (vgl. Steiner 2001, p. 199).

Die Fähigkeit, früh erworbenes Wissen auf neue Aufgabenstellungen oder Lernsituationen zu übertragen (Lerntransfer), ist im schulischen Kontext jedoch mit Schwierigkeiten verbunden. Diese bestehen nicht nur darin, die Umstrukturierung und Verknüpfung von bereits erworbenem (deklarativem) Wissen im Unterricht zu fördern, sondern aus diagnostischer Sicht auch in der Generierung von Frage-Items resp. in der Erfassung der vollbrachten Transferleistung. So haben Beobachtungsstudien amerikanischer Lehrpersonen ergeben, dass sie so gut wie nie Hinweise auf Gedächtnisstrategien geben und auch wenig darauf hinarbeiten, damit Schülerinnen und Schüler abstrahiertes Wissen erwerben können. Es muss wesentlich mehr getan werden, den frühen Erwerb von abstrahiertem Wissen bei Schülerinnen und Schülern zu fördern und entsprechende Diagnoseinstrumente zu erarbeiten (vgl. Helmke & Schrader 2003, p. 199).

2.2.3 Leistungsbeurteilung – Verpflichtung, Gegenstand und Form

Schulen sind offene soziale Systeme, die vom Staat kontrolliert und beeinflusst werden. Ein Hauptauftrag der Schule liegt in der Qualifikation; also darin, Begabungen, Interesse und Persönlichkeit der einzelnen Schülerinnen und Schüler zu fördern und Wissen, Kenntnisse und Fähigkeiten zu vermitteln. Eine solche Förderung und Wissensvermittlung setzt voraus, dass sich Lehrpersonen über Leistungsstand, Begabungen, Potenzial, Interessen, Vorwissen, Wissenszuwachs, Lernmotivation, Ängstlichkeit, Attributionsstil usw. der Lernenden ein möglichst adäquates Urteil bilden können. Ohne Vergewisserung über die Lernforschritte der Schülerinnen und Schüler und Beobachtung ihrer Lernschwierigkeiten unterrichtet die Lehrperson an den Schülerinnen und Schülern vorbei (vgl. Arnold & Jürgens 2001, p. 41). Dieser Auftrag der Schule verpflichtet die Lehrpersonen, Leistungen bzw. Lernerfolg von Schülerinnen und Schülern festzustellen, zu bewerten und in geeigneter Form festzuhalten und rückzumelden. Im «Schweizerischen Schulrecht» (Plotke 2003) wird dieser Auftrag wie folgt festgehalten:

> «Die Schule beurteilt nicht allein die intellektuellen, manuellen und physischen Leistungen, sie äussert sich vielfach zur Arbeitshaltung, zum Fleiss, zum Betragen, zur Eingliederung in die Gemeinschaft, ja sogar zur affektiven Entwicklung – kurz zur gesamten Persönlichkeit des Schülers einschliesslich dessen Charakter» (Plotke 2003, p. 423).

Neben der dienstlichen Pflicht, Leistung und Verhalten der Schülerinnen und Schüler zu beurteilen, hat die Lehrperson auch Zeugnisse auszustellen. «Die Gesamtnoten wie auch die Noten für Fleiss, Betragen und dergleichen werden am Ende der Bewertungsperiode im Zeugnis, einem vorgedruckten amtlichen Formular, […] zusammengestellt» (ebd., p. 432). Notengebung und Zeugnisse müssen gemeinsam betrachtet werden, da Zeugnisse die Leistungsrückmeldungen (meist in Form von Noten) eines grösseren Zeitabschnittes zusammenfassen.

Zum Gegenstand der Beurteilung können

> «Hausaufgaben und -arbeiten; schriftliche und praktische Arbeiten in der Klasse; mündliche Leistungen während der Lektion (während des laufenden Unterrichts oder gleich zu Beginn der Lektion zur Kontrolle, wie die Hausaufgaben erledigt worden sind); spezielle Prüfungen in einem besonderen, formalisierten Verfahren (schriftliche, mündliche, praktische Examina) werden.» (ebd., p. 425).

In unserem Schulsystem ist die Klassenarbeit oder Lern(ziel)kontrolle als formalisiertes Verfahren *das* Standardverfahren zur Feststellung von Lernleistungen geworden (vgl. Arnold & Jürgens 2001, p. 71). Die Klassenarbeiten basieren in der Regel auf den im Unterricht behandelten Inhalten. Die Inhalte bzw. die Lernziele «ergeben sich aus den Lehrplänen, aus Stoffabsprachen beispielsweise mit weiterführenden Schulen [und] aus den Lerninhalten der Lehrbücher» (ebd., p. 423f.).

In der Regel erbringen die Schülerinnen und Schüler die Leistung für sich alleine. Je nach Aufgabe werden jedoch auch Gruppenprüfungen und -arbeiten durchgeführt. «Diese müssen so gestaltet sein, dass sich die Leistungen jedes einzelnen Beteiligten erkennen und bewerten lassen» (ebd., p. 425).

Schriftliche und praktische Klassenarbeiten bzw. Prüfungen erlauben, sofern die in der Diagnostik gebräuchlichen Gütekriterien (vgl. Abschnitt 1.4) erfüllt sind, eine einigermassen gleichmässige und faire Beurteilung. Hausaufgaben bergen hingegen die Gefahr unzulässiger Hilfe. Die Validität wäre damit in grossem Masse verletzt, da die gezeigte und beurteilte Leistung vom «wahren Wert» abweichen kann. Dieses Problem verliert jedoch mit zunehmender Ausbildung an Bedeutung, ab Sekundarstufe II sind bewertete Hausaufgaben (z.B. Matura- oder Diplomarbeiten) üblich.

2.2.4 Leistungsfeststellung, -bewertung und -rückmeldung

Professionelles Verhalten von Lehrpersonen zeichnet sich dadurch aus, dass sie zwischen den drei Teilbereichen der Beurteilung, nämlich der Leistungsfeststellung, -bewertung und -rückmeldung, unterscheiden.

In einem ersten Schritt sind Lehrpersonen aufgefordert, die erbrachte Leistung möglichst exakt festzustellen. Noch relativ einfach scheint dies z.B. bei Diktaten (10 Fehler im Diktat) oder in der Mathematik. Das Messergebnis «ergibt sich aus der Korrektur – zum Beispiel sind 17 von 20 Aufgaben richtig gelöst worden» (Criblez 2002, p. 8). Schwieriger wird dies bereits, wenn die Fehler in Diktaten zu gewichten oder bei den Lösungsvorschlägen in Mathematik auch richtige Ansätze bzw. richtig gelöste Teile der Aufgaben festzustellen sind. Letztlich lässt sich nur über die Diagnose der Leistung feststellen, wie es um das Vorwissen oder die Lernergebnisse bzw. den Wissenszuwachs der Schülerinnen und Schüler steht. Die Feststellung der je unterschiedlichen Leistungsniveaus sowie der gemachten Fortschritte sind für die Planung der weiteren Lehr- und Lernprozesse und damit zur Optimierung der individuellen Lernerfolge elementar.[37]

---

[37] Mit didaktischen und methodischen Prinzipien wie Individualisierung, Differenzierung, Motivierung und Kompensation einerseits, Unterrichtsform und -organisation wie Freiarbeit, Wochenplan- und Werkstattunterricht oder Leitprogrammen andererseits, wird versucht, den individuellen Lernwegen von Schülerinnen und Schülern gerecht zu werden (vgl. Dinges 2002). Aber auch bei frontalen Unterrichtsformen sind Lehr-

Zweitens gilt es, die festgestellte Leistung einer Leistungsbewertung zu unterziehen. Es ist zu klären, was die Schülerinnen und Schüler gelernt haben, wer weniger und wer mehr gelernt hat, und wer die vorgegebenen und durch den Unterricht angestrebten Ziele erreicht hat. Um eine Rückmeldung überhaupt zu ermöglichen, muss eine bestimmte Leistung bewertet und in Beziehung zu anderen Grössen gesetzt werden: Handelt es sich um eine überragende, eine durchschnittliche oder eher magere Leistung? Ist ein Fortschritt feststellbar? Inwiefern entspricht die Leistung dem Durchschnitt Gleichaltriger? Wurde das Lernziel erfüllt? Spätestens hier müssen sich Lehrpersonen über die anzuwendende Bezugsnorm (vgl. Abschnitt 1.5) oder Referenz im Klaren sein.

Der Prozess der Leistungsbeurteilung wird mit dem Schritt der Leistungsrückmeldung abgeschlossen. Die erbrachte Leistung könnte nun ausführlich und differenziert anhand textlicher Leistungsbeschreibungen oder Wortgutachten den Schülerinnen, Schülern und Eltern kommuniziert oder zurückgemeldet werden. Die Rückmeldung erfolgt, indem die «korrigierten Prüfungen verteilt und vielleicht besprochen werden» (ebd., p. 8). Anstelle von ausformulierten Rückmeldungen ist es jedoch oft üblicher, die Fülle von Einzelinformationen in einem einzigen Zeichen, einer Ziffernnote von 1 bis 6, zusammenzufassen.

Die Ziffernnote steht also meist am Ende eines Beurteilungsprozesses: Nach der Leistungsfeststellung folgt die Leistungsbewertung, die das Setzen einer Bewertungsnorm voraussetzt. «Der Unterschied zwischen Messung und Beurteilung besteht darin, dass die Messung sich nur auf objektivierbare, genau bestimmte Kriterien beziehen kann und dass sie die gemessene Grösse nicht interpretiert» (ebd., p. 9). Die Bewertung verlangt einen Massstab, anhand dessen die Lehrperson ein Urteil über die erbrachte Leistung fällen kann. Erst im dritten Schritt wird die Leistung meist mittels eines Codes, namentlich der Ziffernnote, zurückgemeldet. Es steht den Lehrpersonen frei, als Mitteilungsform die formale Skala der Ziffernnote oder eine Kombination von Noten und verbaler Beschreibung zu wählen.

2.2.5 Funktionen der schulischen Leistungsbeurteilung

In Abschnitt 2.2.1 wurde aufgezeigt, dass Leistungsbeurteilung zur dienstlichen Pflicht jeder Lehrperson gehört. Wie wir weiter gesehen haben, ist Leistungsbeurteilung in die Bereiche der *Leistungsfeststellung* (die möglichst exakte quantitative Erfassung der Schulleistung), die *Leistungsbewertung* (die Bewertung einer gemessenen Schulleistung anhand expliziter oder impliziter Normen) und die *Leistungsrückmeldung* (das mündliche oder schriftliche Verhalten der Lehrperson nach einer Leistungsfeststellung und -bewertung) zu unterteilen. Entscheidend ist letztlich aber, welche Funktion oder welchen Zweck eine Leistungsbeurteilung verfolgt. Zentral sind die *Förderungs-* und die *Selektionsfunktion*. Sie bilden gleichsam ein Gegensatzpaar mit förderorientierter Beurteilung auf der einen Seite und selektionsorientierter Beurteilung auf der anderen. Es soll bereits an dieser Stelle darauf hingewiesen werden, dass es nicht darum gehen kann, das eine oder das andere als die bessere oder einzig richtige Funktion darzustellen.

Fend (1981) unterscheidet vier traditionelle Funktionen der Schule als Ganze: Qualifikation, Selektion, Legitimation und gesellschaftliche Kontrolle. Die Funktionen Qualifikation und Selektion stehen dabei in engem Bezug zu den genannten Funktionen der Beurteilung.

Die *Qualifikationsfunktion* steht meist im Zentrum einer aus eher wirtschaftlicher Sicht geführten Diskussion um das Bildungssystem. Es ist anzunehmen, dass das Bildungssystem Eigenschaften vermittelt, die für das Berufsleben bedeutsam sind. Neben direkt verwendbaren Fähigkeiten oder Fertigkeiten sind damit auch Internalisierungen von Werten wie Pflichtbewusstsein, Pünktlichkeit und Fleiss gemeint. Um diese Qualifikationsfunktion erfüllen zu können, muss die Beurteilung ein

---

personen auf Leistungsdiagnostik angewiesen, sie kann ihnen z.B. als Instrument bei der Wahl der Inhalte, den Schwierigkeitsgraden und der Organisationsform dienen.

Element der Unterrichtsplanung und -gestaltung sein. Beurteilung ist in diesem Sinne eine Rückmeldung an Schülerinnen, Schüler und Lehrpersonen; dank ihr kann der weitere Lernprozess gesteuert und die Qualifikation optimiert werden. Folgende förderorientierte Absichten sind mit der didaktischen Kontrolle des Lernerfolgs verbunden (vgl. Rhyn & Moser 2002, p. 33):

- Korrektur falscher Lernergebnisse;
- Ausgleich von Lerndefiziten;
- Bestätigung erfolgreicher Lernschritte;
- Planung der folgenden Lernschritte;
- Motivierung durch Steuerung der Aufgabenschwierigkeiten;
- Optimierung der Lernbedingungen.

Neben der *Qualifikation* ist auch *Selektion* im Bildungssystem unabdingbar. Seit in den Kantonen der Schweiz nach 1830 moderne Rechtsstaaten mit entsprechenden Bildungssystemen aufgebaut worden sind, in denen die früheren Vorrechte von Geburt und Stand abgeschafft wurden, hat sich das Leistungsprinzip als Ordnungsprinzip der modernen Gesellschaft durchgesetzt (vgl. Criblez 2002, p. 83). Das Bildungssystem liefert seinen Beitrag dazu, indem die Grundannahmen des Prinzips der Chancengleichheit[38] bei der Beurteilung und der darauf basierenden Selektion in unterschiedliche Schultypen bzw. weiterführende Schulen berücksichtigt werden. Wichtigste Grundannahmen dabei sind: a) Allein die erbrachte Leistung bzw. der Lernerfolg entscheiden bei der Auslese und Zuteilung zu höheren Schulstufen. b) Der Erfolg liegt in der Hand des Individuums. Es liegt an ihm, sich entsprechend seinen Fähigkeiten einzusetzen und zu bewähren. c) Niemand darf wegen seiner sozialen Herkunft, seines Geschlechtes, seines Wohnortes, seiner Sprache oder seiner religiösen oder ethnischen Zugehörigkeit im Bildungssystem benachteiligt werden (vgl. Lamprecht & Stamm 1996).

Aufgrund der Selektionsentscheidungen werden die Schülerinnen und Schüler in Schultypen mit unterschiedlichen Anforderungsniveaus eingeteilt. Es ist jedoch mit Überschneidungen zu rechnen, dass z.B. die besten Schülerinnen und Schüler von Schultypen mit niederen Anforderungen bessere Leistungen erbringen als schlechte Schülerinnen und Schüler von Schultypen mit höheren Anspruchsniveaus (vgl. Abbildung 2.1).

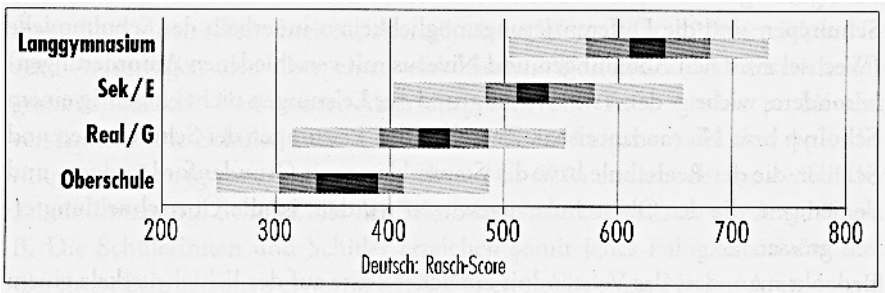

Abbildung 2.1: Verteilung der Deutschleistungen nach dem Übertritt der Schülerinnen und Schüler in die Schultypen der Sekundarstufe I des Kantons Zürich (aus: Moser & Rhyn 2000, p. 41)

Daraus ergibt sich unmittelbar die Forderung, dass Differenzierungsmöglichkeiten (Wechsel der Abteilungen bzw. Niveaus) und/oder Durchlässigkeiten auch nach den Selektionsentscheidungen möglich sein sollen. Frühe Selektionsentscheide dürfen nicht lebenslänglich wirken. Ditton (1992)

---

[38] Ob diese Chancengleichheit im schweizerischen Bildungssystem tatsächlich gegeben ist, beispielsweise im Falle des Schulerfolgs und der Bildungsbeteiligung von Kindern und Jugendlichen mit Migrationshintergrund (oder in Bezug auf Geschlecht und soziale Schichtzugehörigkeit), ist eine Frage, welche u.a. seit den Resultaten von PISA mehr denn je gestellt und diskutiert wird (vgl. dazu Kronig, Haeberlin & Eckhart 2000; Coradi Vellacott 2003).

weist jedoch zu Recht darauf hin, dass derartige Sprünge in der Praxis nur sehr begrenzt stattfinden und dass einmal selektionierte Schülerinnen und Schüler gleichsam wie festgefroren in der zugewiesenen Stufe verbleiben:

> «Im Zuge der Bildungsexpansion haben die Raten des Versagens in den höheren Schulen abgenommen. […] Auf der anderen Seite ist die Durchlässigkeit nach oben, der nachträgliche Aufstieg in höherwertige Schullaufbahnen, sehr begrenzt. [… Die Aufstiegschancen sind] sehr gering, und das auch dann, wenn die Leistungen und Leistungsbedingungen ursprünglich gut waren, gemessen im Vergleich zu Schülern, die mit teils weitaus schlechteren Schulleistungen eine höherwertige Bildungslaufbahn eingeschlagen haben» (Ditton 1992, p. 72).

Damit verschärft sich die Situation, und die Selektion wird für die weitere Bildungslaufbahn noch wichtiger und entscheidender.

Neben der vorgenommenen Unterteilung der Leistungsbeurteilung in Förder- und Selektionsorientierung sind zwei weitere Funktionen der Notengebung nicht zu vergessen: Jede Messung der je individuellen Lernerfolge der Schülerinnen und Schüler liefert bis zu einem bestimmten Grade auch Anhaltspunkte über die Wirksamkeit des erteilten Unterrichts. Die Lehrperson ist «in der schwierigen Lage, sozusagen gleichzeitig auch ihren eigenen Erfolg messen und bewerten zu müssen» (Konradt 1975, p. 124). Die Beurteilung der Leistungen der Schülerinnen und Schüler dient deshalb auch der *Selbstevaluation* für Lehrpersonen. Leistungsbeurteilung kann zur Standortbestimmung, Überprüfung und Optimierung des eigenen Unterrichts dienen (vgl. Moser & Rhyn 2000, p. 137). Die Fähigkeit zur Unterrichtsplanung, -vorbereitung und -durchführung kennzeichnen im Wesentlichen die Professionalität der Lehrpersonen. Werden diese Phasen mit der Unterrichtsauswertung anhand der Leistungen der Schülerinnen und Schüler erweitert, kann von einer Selbstevaluation oder Evaluation des Lehr-Erfolges gesprochen werden. Letztlich ist es aber nicht möglich, die Wirkung des Unterrichts von der Wirkung, die vom Individuum und seiner je spezifischen Umwelt ausgeht, zu trennen.[39] Als weitere Funktion der Leistungsbeurteilung geht man «in der Schule häufig davon aus, dass gute Noten als Bestätigung und Belohnung empfunden werden, und dass eine schlechte Note Ansporn oder wenigstens Aufforderung zu grösseren Anstrengungen ist» (Konradt 1975, p. 123). Konradt nennt dies die *erzieherische Funktion* von Leistungsbeurteilung. Berücksichtigt eine Lehrperson diese Funktion bei der Leistungsfeststellung, so wirkt sich dies stark auf die Objektivität und Vergleichbarkeit der einzelnen Note aus. Der gedachte Zusammenhang ist aber nicht so eng, wie man ihn sich idealerweise vorstellt. Zu viele Motivationsprozesse beeinflussen die erwünschte motivierende Steuerung. Es besteht die Gefahr, dass die Note wegen ihrer vermeintlichen erzieherischen Wirkung von Lehrpersonen als Disziplinierungsmassnahme eingesetzt wird: Unangekündigte und damit für Schülerinnen und Schüler unerwartete Klassenarbeiten sollen die Aufmerksamkeit und Teilnahme im Unterricht erhöhen.

2.3 Diagnosekompetenz von Lehrpersonen

Die diagnostischen Fähigkeiten der Lehrpersonen stellen eine wichtige Kompetenz im Unterricht dar. Sie sind neben der Klassenführungskompetenz, der didaktischen und der fachwissenschaftlichen Kompetenz eine der vier Schlüsselkompetenzen und somit «unabdingbare Voraussetzung für effektiven Unterricht im allgemeinen und individuelle Förderung im besonderen» (Helmke, Hosen-

[39] Metaanalysen zeigen zumindest, dass die Qualität des Unterrichts zusammen mit den Merkmalen der Schülerinnen und Schüler sowie ausserschulische Einflüsse die Hauptverantwortung für die schulischen Leistungen tragen. Lehrpersonen sind daher kaum in der Lage, aufgrund der schulischen Leistungsdiagnostik den Eigenanteil des Lernfortschrittes der Schülerinnen und Schüler genau zu quantifizieren. Zu vielfältig und vernetzt sind die einzelnen Einflussfaktoren. Eine misslungene Klassenarbeit oder eine Aufgabenstellung, die nur von wenigen gelöst werden konnte, kann der Lehrperson jedoch als ein Indiz dazu dienen, dass eventuell die Qualität des eigenen Unterrichts zu überdenken und die eigene Arbeit zu optimieren ist.

feld & Schrader 2004, p. 119). Zu Beginn dieses Abschnitts sollen einige wichtige Begriffe der Diagnosekompetenz von Lehrpersonen erläutert werden. In einem zweiten Teil werden empirische Ergebnisse zur diagnostischen Kompetenz von Lehrpersonen vorgestellt.

2.3.1 Was ist Diagnosekompetenz?

Der Gegenstandsbereich der Diagnosekompetenz lässt sich laut Helmke et al. (2004) in zwei Bereiche einteilen. Zum einen wird unter Diagnosekompetenz von Lehrpersonen häufig die Fähigkeit verstanden, Schülerinnen und Schüler zutreffend einzuschätzen (personenbezogene Diagnosefähigkeit). Sie wird somit mehr oder weniger mit Diagnose- oder Urteilsgenauigkeit gleichgesetzt. Grundlage dafür sind meistens Lehrerinnen- und Lehrerurteile über die Leistungen der Lernenden, die mit den tatsächlich von den Schülerinnen und Schülern erzielten Leistungen verglichen werden. Der Grad an Übereinstimmung zwischen Lehrerinnen- und Lehrerurteilen und den Leistungen der Lernenden wird dann als Indikator für die diagnostische Kompetenz der urteilenden Lehrperson gesehen. Andererseits wird unter der Diagnosekompetenz auch die Genauigkeit der Beurteilung von Aufgabenschwierigkeiten untersucht. Lehrpersonen schätzen dazu vorgängig den Schwierigkeitsgrad verschiedener Testaufgaben ein, die dann mit deren tatsächlichem Schwierigkeitsgrad (Prozentsatz der Schülerinnen und Schüler, die die Aufgabe lösen konnten) verglichen wird. Leitender Gesichtspunkt der diagnostischen Kompetenz ist die Adaptivität, «d.h. die Notwendigkeit einer Anpassung des Unterrichts an die Lernvoraussetzungen der Schülerinnen und Schüler (Prinzip der Passung)» (ebd., p. 8).

Grundlage für die Fähigkeit, genaue diagnostische Urteile abgeben zu können, sind nach Helmke et al. (2004, p. 9f.) stabile und schwer beeinflussbare kognitive Merkmale der Lehrperson (z.B. Intelligenz), bereichsspezifische Fähigkeiten und Wissensstrukturen (methodisches Wissen, gegenstandsbezogenes Wissen) und spezifische Kenntnisse (Wissen über einzelne Schülerinnen und Schüler oder Wissen über die Klasse). Das Wissen und die Forschung über diagnostische Kompetenzen haben im deutschsprachigen Raum eine geringe Verankerung, während ihnen in den USA traditionell eine grosse Wichtigkeit beigemessen wird. Dazu exemplarisch die Standards zur diagnostischen Kompetenz des amerikanischen Lehrerverbands (vgl. Tabelle 2.1):

Tabelle 2.1: Standards des amerikanischen Lehrerverbands zur diagnostischen Kompetenz von Lehrpersonen (aus: Arnold 1999)

1. Lehrkräfte sollen aus dem verfügbaren Methodenarsenal jene Diagnoseverfahren auswählen können, die für anstehende Unterrichtsentscheidungen angemessen sind.
2. Lehrkräfte sollen Diagnoseverfahren entwickeln können, die für anstehende Unterrichtsentscheidungen angemessen sind.
3. Lehrkräfte sollen fähig sein, sowohl von Testexperten professionell konstruierte als auch unterrichtsbezogene, von Lehrkräften entwickelte Diagnoseverfahren anzuwenden, auszuwerten und zu interpretieren.
4. Lehrkräfte sollen fähig sein, pädagogisch-diagnostische Informationen (a) in pädagogischen Entscheidungen, die einzelne Schülerinnen und Schüler betreffen, (b) für die Planung ihres Unterrichts, (c) im Rahmen der Curriculumsentwicklung und (d) in Schulentwicklungsprozessen zu nutzen.
5. Lehrkräfte sollen fähig sein, valide Verfahren der zusammenfassenden Beurteilung von Einzelbewertungen zu entwickeln.
6. Diagnoseergebnisse sollen von Lehrkräften in verständlicher Weise den Schülerinnen und Schülern, Eltern oder anderen Laienpersonen mitgeteilt werden können.
7. Lehrkräfte sollen fähig sein, ethisch inakzeptable oder ungesetzliche oder in anderer Weise unangemessene Diagnoseverfahren bzw. eine in dieser Weise problematische Anwendung von Verfahren zu erkennen.

Helmke et al. (2004, p. 4f.) unterscheiden drei unterschiedliche Komponenten der Diagnosegenauigkeit: Die *Niveaukomponente* bezieht sich darauf, ob die Lehrperson im Mittel das zur Debatte stehende Merkmal des Schülers resp. der Schülerin als zu hoch, gerade richtig oder zu niedrig einschätzt. Die *Streuungskomponente* bezieht sich auf den Vergleich der durch die Lehrpersonen geschätzten Streuung zu der tatsächlich von Schülerinnen und Schülern in einem Test erzielten Streuung[40]. Das «Kernstück der diagnostischen Kompetenz» (ebd.) ist die *Rangordnungs- oder Korrelationskomponente*. «Sie betrifft die Fähigkeit, die Rangordnung bzw. Fähigkeitsabstufung zwischen verschiedenen Schüler/-innen (oder Aufgaben) zu erkennen» (ebd.).

Für die Qualität von Lehrerinnen- und Lehrerurteilen gelten zudem die gleichen Kriterien wie in der Testtheorie (vgl. Abschnitt 1.3 und 1.4). Die Gütekriterien der Qualität von Lehrerurteilen sind demnach Objektivität (inwiefern stimmen Lehrpersonen in ihrem Urteil überein, wenn sie den gleichen Sachverhalt messen), Reliabilität (Genauigkeit, mit der ein Messinstrument ein Merkmal misst) und Validität (Ausmass, in dem Lehrpersonen das beurteilen, was beurteilt werden soll). Hinsichtlich der angewendeten Bezugsnormen bei der Bestimmung des Standorts einer Klasse, wird die begrenzte Reichweite diagnostischer Urteile im Schulalltag deutlich. Laut Schrader (2001) orientieren sich Lehrpersonen primär an einem klasseninternen Bezugssystem, da ihnen für die klassenübergreifende Beurteilung oft die nötigen Vergleichsinformationen fehlen und solche Urteile nicht Teil der alltäglichen Diagnosepraxis sind.

2.3.2 Empirischer Forschungsstand

Laut Helmke (2003, p. 93) wurde der diagnostischen Kompetenz von Lehrpersonen seit den frühen Arbeiten von Gage (1968) nur wenig Aufmerksamkeit geschenkt. Neben einigen neueren Studien aus Amerika erwähnt Helmke die von Weinert initiierten Arbeiten (Schrader 1989; Weinert, Schrader & Helmke 1990; Weinert & Lingelbach 1995) sowie ein eigenes Forschungsprojekt (Hosenfeld, Helmke & Schrader 2002). Ingenkamp leistete im Bereich der Erforschung schulischer Leistungsdiagnostik Pionierarbeit. Er gab eine Reihe von Untersuchungen unter dem Titel «Die Fragwürdigkeit der Zensurengebung» (1971) heraus und machte auf die Mängel der Zensurenpraxis aufmerksam. Die Publikation ging inzwischen in die zehnte Auflage und die Ergebnisse haben, so Jachmann (2003, p. 45), nicht an Aktualität und Brisanz verloren. So ist laut Helmke et al. (2004) die bisherige Forschung zur diagnostischen Kompetenz von Lehrpersonen wenig umfangreich, und sie beschränkt sich weitgehend auf einen Teilaspekt der diagnostischen Kompetenz, nämlich auf die Diagnosegenauigkeit bei kognitiven Leistungen. In den meisten Untersuchungen werden klassenspezifische Korrelationen zwischen Lehrerurteilen und korrespondierenden Schülerleistungen in Bezug auf die Rangordnung dargestellt. Dabei hat sich gezeigt, dass sich die mittleren Korrelationen meistens zwischen $r = .60$ und $r = .70$ bewegen. Daraus wird häufig geschlossen, dass Lehrpersonen recht gut über die Leistungsunterschiede in ihrer Klasse Bescheid wissen (vgl. ebd., p. 7).

Helmke et al. (ebd.) weisen allerdings darauf hin, dass das Prinzip «je genauer, desto besser» für Lehrerurteile zu relativieren ist. Einiges spricht dafür, dass eine leicht positiv verzerrte Überschätzung des Leistungspotentials der Schülerinnen und Schüler *pädagogisch* günstig ist. Eine derartige Beurteilungspraxis führt dazu, dass den Schülerinnen und Schülern anspruchsvolle Aufgaben gestellt und ihnen so günstige Entwicklungsanreize geboten werden (vgl. Schrader 2001). Gemäss Helmke et al. (2004) wäre es ausserdem weltfremd, wenn man von Lehrpersonen extrem präzise diagnostische Leistungen verlangen würde. Die Unschärfe des Wissens bei Fähigkeitsunterschieden von Schülerinnen und Schülern muss in einem Bereich liegen, der tolerierbar ist.

---

[40] Streuung meint die «Breite» des Ausprägungsbereiches, in den alle Merkmalsausprägungen fallen. Sie wird auch als Variabilität bezeichnet: V = (maximaler Ausprägungsgrad) – (minimaler Ausprägungsgrad des Merkmals) (vgl. Hirsig 2001, p. 240).

Über die Kompetenz bei der Einschätzung von Aufgabenschwierigkeiten liegen nur sehr wenige Untersuchungen vor. In ihnen zeigte sich aber eine Tendenz zur Überschätzung des Niveaus der Klassenleistung (vgl. Helmke et al. 2004, p. 8). Dies sei exemplarisch an der Lehr-Lern-Studie (SALVE) von Hosenfeld, Helmke und Schrader (2002) dargestellt (vgl. Abbildung 2.2): Die Einschätzung der Aufgabenschwierigkeit zeigt, dass die Lehrpersonen durchschnittlich von einem Anteil von 58,7% der Schülerinnen und Schüler ausgehen, welche die zu beurteilenden 12 Aufgaben lösen konnten. Die Spanne der Einschätzungen reicht dabei von 35,7% bis zu einem Maximum von 82,8%. Die empirischen Lösungshäufigkeiten der Klassen liegen – mit Ausnahme einer Klasse – unter den Schätzungen der Lehrpersonen. Im Durchschnitt überschätzen die Lehrpersonen ihre Klassen um 18%, das Maximum liegt hier bei 45,5% (vgl. ebd. p. 72f.).

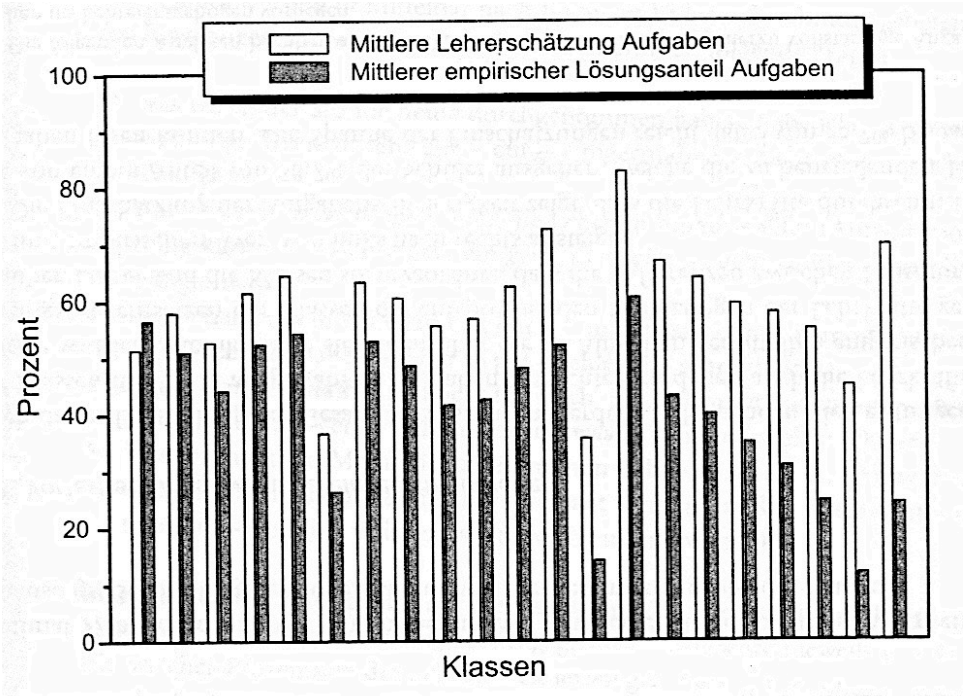

Abbildung 2.2: Mittlere Einschätzung der Lehrpersonen hinsichtlich der Aufgabenschwierigkeit und mittlere empirische Lösungshäufigkeit für 21 Klassen (aus: Hosenfeld, Helmke & Schrader 2002, p. 73)

Zur unterschiedlichen Verteilung von Noten zitiert Jachmann (2003, p. 53) eine Hamburger Studie, in der die Hypothese bestätigt wurde, dass Mädchen eher bessere Noten erhalten als Jungen, auch wenn die Testleistungen dies nicht erwarten lassen. Es zeigte sich, dass Kinder aus bildungsferneren Schichten bei der Zensurierung tendenziell benachteiligt werden. Weiter wurde bei der Analyse der Verteilung von Noten deutlich, dass nicht alle Fächer gleich streng bewertet werden. Sogenannte Haupt- oder Leistungsfächer werden strenger beurteilt als z.B. Gestaltungsfächer.

Angesichts der eingangs erwähnten Bedeutung der diagnostischen Kompetenz von Lehrpersonen ergibt sich die Frage, wie man sie fördern kann. Wahl, Weinert und Huber (1997) empfehlen Lehrpersonen, möglichst häufig Situationen herzustellen, in denen Leistungen von Schülerinnen und Schülern vorhergesagt und dann erfasst werden können. Die Vorhersagedaten werden dann mit den tatsächlichen Leistungen, den Testdaten, verglichen. Das Grundprinzip dieses Ansatzes liegt darin, im Schulalltag oft nur implizit abgegebene Urteile explizit zu machen und sie dann mit den tatsächlichen Ergebnissen zu vergleichen. Damit wird eine Reflektion der zugrunde liegenden Wissensbasis angeregt. Diskrepanzen zwischen erwarteten und tatsächlich erzielten Leistungen können dazu genutzt werden, um nach Gründen für die erwartungswidrigen Ergebnisse zu suchen.

Helmke et al. (2004) beschreiben eine Lehrperson dann als kompetenten Diagnostiker resp. kompetente Diagnostikerin, wenn sie sich auf spezifische und aktuelle Beobachtungen stützt und sich um präzise und überprüfbare Vorhersagen bemüht und sich nicht von allgemeinen und vagen Erwartungen und Einschätzungen leiten lässt. Wichtig ist ein solides, handlungsnahes Wissen über Urteilsbildung, Urteilsfehler und diagnostische Methoden sowie eine professionelle diagnostische Einstellung, die dadurch gekennzeichnet ist, dass eine ständige Überwachung und Überprüfung der eigenen Diagnoseleistungen als wichtiger Bestandteil der Berufsrolle betrachtet wird (vgl. ebd.).

Laut Helmke et al. (2004) ist diagnostische Kompetenz vor allem deshalb von Interesse, weil ihr eine wichtige Rolle für die Unterrichtsgestaltung und den Unterrichtserfolg zugeschrieben wird. Der Unterricht muss an die Lernvoraussetzungen der Schülerinnen und Schüler angepasst werden. Dafür ist es wichtig, dass Lehrpersonen hinreichend über den Leistungsstand, die verfügbaren Vorkenntnisse und das erzielte Lernergebnis orientiert sind. In ihren Untersuchungen zeigte sich aber, dass hohe diagnostische Kompetenz nur dann für den Lernerfolg der Schülerinnen und Schüler bedeutsam war, wenn sie mit einem hohen Mass an Strukturierungshilfen einerseits und einem hohen Mass an individueller fachlicher Unterstützung andererseits einherging. Der Lernerfolg war besonders hoch, wenn Lehrpersonen mit einer hohen Diagnosekompetenz gleichzeitig viele Strukturierungs- und Unterstützungsmassnahmen einsetzten. Die blosse Fähigkeit, Leistungen korrekt vorherzusagen, reicht nicht aus, es müssen demnach auch Schritte getroffen werden, die die Ergebnisse der Leistungsmessung in der Schule für die weiteren Lernprozesse der Schülerinnen und Schüler nutzbar machen. Eine ähnliche *Wechselwirkung zwischen diagnostischer Kompetenz und Unterrichtskompetenz* konnte auch in einer Folgestudie von Weinert und Lingelbach (1995) nachgewiesen werden.

In den folgenden Abschnitten geht es nun um unterschiedliche Instrumente, die von Lehrpersonen im Rahmen ihrer diagnostischen Tätigkeit eingesetzt werden können. Dabei wird zwischen nicht-standardisierten und standardisierten Verfahren unterschieden, wobei der Notengebung besondere Aufmerksamkeit geschenkt wird.

2.4 Nicht-standardisierte Verfahren zur schulischen Leistungsdiagnostik

Nicht-standardisierte Schulleistungstests sind Verfahren der pädagogischen Diagnostik, die durch Lehrpersonen konstruiert werden, um die Ergebnisse der von ihnen geplanten Lernvorgänge in ihrer Klasse möglichst objektiv zu erfassen und für ihr pädagogisches Handeln nutzbar zu machen (vgl. Ingenkamp 1997, p. 131). Das wichtigste Unterscheidungsmerkmal zu den standardisierten Tests besteht darin, dass die Konstruktion des Tests durch die Lehrperson vorgenommen wird. Ein anderes Unterscheidungsmerkmal gegenüber standardisierten Tests ist die Orientierung an den pädagogischen Bedürfnissen *einer* (oder weniger) Schulklassen. Unter den Begriff des nicht-standardisierten Tests fallen sowohl die gängigen Klassenarbeiten als auch stärker analysierende Methoden, wie sie von Ingenkamp (1997) und Lukesch (1998) dargestellt werden. Dabei werden bestimmte Phasen durchlaufen[41]:

In der *Planungsphase* stehen die Zielsetzungen des eigenen Unterrichts und methodische Entscheidungen an: was, warum und wie soll gemessen werden? In der Folge orientiert sich die Lehrperson bei der *curricularen Analyse* an den Lehrzielen, die sie für ihre Klasse ausgewählt hat. Bei der *Aufgabenkonstruktion* müssen Lehrpersonen entscheiden, welche Aufgabentypen den Zielsetzungen, Inhalten und Anwendungsmöglichkeiten am besten entsprechen. Ingenkamp (1997) empfiehlt

[41] Die verschiedenen Phasen der Testkonstruktion finden sich ähnlich wieder im Abschnitt 2.6 Standardisierte Schulleistungstests.

Lehrpersonen mit geringer Erfahrung ihre Entwürfe mit erfahrenen Kollegen und Kolleginnen zu diskutieren (ebd. p. 132). Die empirische Aufgabenanalyse wird nach vereinfachten Verfahren durchgeführt. Hier können statistische Kennwerte wie Aufgabenschwierigkeit[42] und Trennschärfe[43] berechnet werden. Eine Eichung des Tests an einer repräsentativen Vergleichsgruppe findet nicht statt.

> «Informelle Tests [resp. nicht-standardisierte Tests] können dazu beitragen, dass die Lehrer zur gründlicheren curricularen Planung angehalten werden und dass viele der traditionellen Kurzarbeiten und mündlichen Prüfungen durch eine objektivere Methode ersetzt werden» (Ingenkamp 1997, p. 134).

Für die individuellen Bedürfnisse der Lehrpersonen wurden nicht-standardisierte Schulleistungstests mit geringerem methodischen Aufwand entwickelt. Nicht-standarisierte Tests tragen dazu bei, dass viele traditionelle Arbeiten durch eine objektivere Methode ersetzt werden. Nach Ingenkamp (1997, p. 133) spielen jedoch die Phasen Testkonstruktion, Durchführung und statistische Auswertung von nicht-standardisierten Tests in unserem Schulwesen eine untergeordnete Rolle.

Der Begriff des nicht-standardisierten Schulleistungstests kann auch weiter gefasst werden. So beschreibt er alle Tests, die von den Lehrpersonen geschaffen werden. Unterschiede zeigen sich im Grad der Standardisierung resp. in der Bedeutung, die den oben beschriebenen Phasen zugemessen wird.

Im Folgenden werden einige nicht-standardisierte Testverfahren dargestellt, die für die schulische Praxis bedeutungsvoll sind.

2.4.1 Mündliche Prüfungen

Die Beurteilung der Lernleistung auf der Grundlage eines Gesprächs gehört zu den ältesten Formen der Lernkontrolle. Sie tauchten in den Schulen des Mittelalters vorwiegend in der Form der Disputation auf. Im heutigen Bildungssystem sind mündliche Prüfungen weit verbreitet, gleichzeitig aber sind sie die am wenigsten erforschte Form der Leistungskontrolle. Charakteristisch für mündliche Prüfungen ist ihr sozialer Aspekt:

> «Mündliche Prüfungen stellen unter einem diagnostischen Gesichtspunkt insofern eine Sondersituation dar, als sie im Rahmen eines sozialen Interaktionsprozesses zwischen Lehrer und Schüler erfolgen und somit verschiedenen Einflüssen der sozialen Wahrnehmung unterliegen» (Wild & Krapp 2001, p. 539).

In diagnostischer Hinsicht dienen mündliche Prüfungen der (abschliessenden) Kontrolle des Lernerfolgs und der Zwischendiagnose von Teilergebnissen im Unterrichtsverlauf (vgl. ebd.). Alle Prüfungen, die sich nicht der schriftlichen Form bedienen, wie Prüfungen im Sportunterricht, Vorträge auf Musikinstrumenten, Gedichtrezitationen, Fahrprüfungen usw. können den mündlichen Prüfungen zugeordnet werden.

Typisch für die schulische Situation sind Gespräche und Befragungen. Alle Modi von Gespräch und Befragung, beginnend mit dem pädagogischen und diagnostischen Gespräch, können in der mündlichen Prüfung auftreten. Mündliche Prüfungen sind methodischen Gesprächen aber übergeordnet. Sie sind mehr oder weniger offen, jedoch strukturiert und laut Definition asymmetrisch (vgl. Flammer 1997, p. 216). Abbildung 2.3 gibt einen Überblick über die verschiedenen Formen von Gespräch und Befragung.

---

[42] Verhältnis von «Anzahl Personen mit richtigen Lösungen» zu «Anzahl Personen, die die Aufgabe bearbeitet haben» (vgl. Ingenkamp 1997, p. 122).

[43] Ein hoher Trennschärfekoeffizient besagt, dass die Personen, welche das Item richtig gelöst haben, auch einen hohen Gesamtwert erhalten (vgl. Lukesch 1998, p. 508).

Kategorien	Modi der Gliederung		
Methode	nicht methodisches Gespräch	methodisches Gespräch / Befragung	
Intention a)	Konversation	pädagogisches bzw. therapeutisches Gespräch	diagnostisches Gespräch
b)	Gelegenheitserkundung	Exploration	Erkundungsgespräch Anamnese "mündliche Prüfung"
Kommunikation	Small talk	Beziehungsebene - Selbstoffenbarung - Sachebene - Appell das "echte Gespräch" (Dialog oder gemeinsame Erörterung) Diskurs symmetrisch ←----------→ asymmetrisch	
Strukturiertheit	unstrukturiert	niedrig strukturiert ←----------→ hoch strukturiert offenes, qualitatives, halbstandardisiertes, standardisiertes Interview	
Objektivierung	nicht protokolliert	fremd-protokolliert ←-----→ selbst-protokolliert Tonträger	Fragebogen, "Test", "schriftliche Arbeiten"
Direktheit	indirekt ←----------→ direkt		
Technik	weiche Technik ←----------→ neutrale Technik ←----------→ harte Technik		

Abbildung 2.3: Gliederung von Gespräch und Befragung unter pädagogisch-diagnostischem Aspekt (aus: Kleber 1992, p. 208.)

In der einschlägigen Literatur werden drei Gesprächstechniken unterschieden (zuunterst in der Abbildung): die weiche, die harte und die neutrale Gesprächs- bzw. Befragungstechnik (vgl. Kleber 1995, p. 224). Eine druck- und zwangfreie Gesprächssituation anhand der humanistischen Psychologie von Rogers ist das Grundgerüst der *weichen* Gesprächstechnik (vgl. ebd.). Die Prüfer und Prüferinnen sollen dabei «Achtung, Wärme, gleich bleibende Freundlichkeit, Offenheit, aktives, offenes Zuhören zeigen und druck- und zwangfreie Gesprächssituationen schaffen» (ebd., p. 225). Die *harte* Gesprächstechnik zeichnet sich vor allem dadurch aus, dass erst durch Konfrontation und aggressive Provokation relevante Aussagen hervorgebracht werden. Das Verfahren ist auch als *Stress-Interview* bekannt (vgl. ebd.). In der *neutralen* Gesprächssituation übernimmt die befragende Person die Rolle einer anonymen Mittelsperson. Die befragende Person verhält sich dabei «höflich, distanziert, sachlich, in jedem Fall unpersönlich» (ebd.). Im pädagogischen Raum sollte nach Kleber (ebd.) die harte Technik vermieden werden. Prüfungssituationen werden vor allem im Sinne der neutralen Technik realisiert.

Alle Modi von Gespräch und Befragung sind mit Problemen behaftet. So können nichtinhaltliche Merkmale (z.B. Stress, Angst, Krankheit) die Prüfungssituation erheblich verzerren. Wie bei jeder anderen diagnostischen Methode hängt die Qualität des Verfahrens von der fachlichen Kompetenz des Beurteilenden, von der Art der Vorbereitung und der bewussten Kontrolle von Störgrössen ab (vgl. Wild & Krapp 2001, p. 539). Die Qualitätsunterschiede zwischen verschiedenen Prüfungssituationen sind beträchtlich. Nimmt man z.B. die Urteilsübereinstimmung verschiedener Prüferinnen und Prüfer (bei den gleichen Kandidaten) als Qualitätskriterium, so variiert diese Interraterreliabilität von Untersuchung zu Untersuchung sehr stark. Ingenkamp (1997, p. 103) berichtet von Untersuchungen, deren Ergebnisse mit $r = .40$ und $r = .60$ eher tief liegen. Ähnliche Unterschiede findet man, wenn Prüflinge innerhalb kurzer Zeit von zwei unabhängigen Prüfungskommissionen im selben Fach geprüft werden (vgl. Wild & Krapp 2001, p. 539).

Es gibt verschiedene Möglichkeiten, um die Qualität einer mündlichen Befragung zu erhöhen. So sollte etwa der Gesprächsverlauf so weit strukturiert sein, dass die Prüfung nicht völlig vom Geschick des Kandidaten resp. der Kandidatin abhängt. Zudem ist es von Vorteil, wenn dem Prüfling die Beurteilungskriterien bekannt sind (vgl. Kleber 1995, p. 226ff.; Flammer 1997, p. 225ff.). Weiter müssen subjektive Fehlerquellen vermieden werden, wie zum Beispiel die Tendenz, keine extremen Noten zu schreiben und die ersten und letzten Prüfungen überzubewerten (vgl. Abschnitt 2.5.1.1).

2.4.2 Aufsatzprüfung

Bei einer Aufsatzprüfung, als typische Form der schriftlichen Prüfung, wird nur die Frage bzw. das Thema vorgegeben (vgl. Wild & Krapp 2001, p. 539f.). Die Schülerinnen und Schüler müssen ihre Antwort frei formulieren und zu Papier bringen. Da die Antworten stark variieren, sind sie nicht unmittelbar vergleichbar. Die individuellen Aussagen eines Prüflings sind offen für verschiedene Interpretationen der Prüfenden und die Beurteilungsmassstäbe sind normalerweise nicht standardisiert. Deshalb sind Urteile über Aufsatzprüfungen sehr selten im strengen Sinne objektiv.

Als Faktoren, die die Beurteilung einer schriftlichen Arbeit beeinflussen, werden genannt (vgl. Ingenkamp 1997, p. 109):

- die Länge der Textproduktion;
- grammatikalische und orthographische Fehler, selbst wenn nur der Inhalt beurteilt wird;
- die Handschrift;
- die Reihenfolge, in der die Arbeiten beurteilt wurden;
- das Geschlecht der Beurteilten und des Beurteilers;
- die Beliebtheit des Schülers resp. der Schülerin.

Wenn die Themen sorgfältig ausgewählt werden und die Auswertung ebenso vorbereitet und strukturiert ist, d.h. zum Beispiel Kriterien im vornherein bestimmt werden, lassen sich grobe Fehler vermeiden. Bewährt haben sich z.B. Punktesysteme, die nach einem festgelegten Schema einzelne Aussagen und Argumentationsstrukturen mit einem Anforderungsprofil vergleichen und je nach Annäherungsgrad die Ideallösung gewichten (vgl. Wild & Krapp 2001, p. 539f.).

Versuche zur Verbesserung der Aufsatzbeurteilung wurden von Beck (1979, zit. n. Ingenkamp 1997, p. 114) wie folgt untergliedert:

- textganzheitlich orientierte Vorgehensweisen: bspw. Teamauswertung (mehrere Lehrpersonen lesen den Aufsatz unabhängig voneinander durch und geben Noten nach dem Gesamteindruck. Der Mittelwert dieser Noten gibt die Schlussnote ab); Einbeziehung von Vergleichsarbeiten (mit skalierten Musteraufsätzen);
- testanalytisch bestimmte Vorgehensweisen: z.B. lernzielbezogene Verfahren (Lernziele werden operationalisiert und daraus Beurteilungskriterien abgeleitet); Einsatz von Kriterienkatalogen;
- kombinierte Verfahren: bspw. Diagnosebogen (vgl. ebd., p. 135ff.).

2.4.3 Selbstbeurteilung

Wenn Schülerinnen und Schüler selber am Beurteilungsprozess teilnehmen, wird von Selbstbeurteilung gesprochen. Diese Selbstbeurteilung beginnt laut Kleber (1995, p. 282) bei der Aufgabe des Inventarisierens auf einem sehr einfachen Niveau, meistens bei der Selbstkontrolle der Aufgabenerledigung oder bei der Bearbeitung des Arbeitsmaterials. Kleber zählt verschiedene Formen der Selbstbeurteilung auf (vgl. ebd., p. 283):

- Lernweg-Leit-Kartei (Selbstkontrolle der richtig erledigten Aufgaben): Die Schülerin oder der Schüler beurteilt aufgrund seiner Lernweg-Leit-Kartei, was er oder sie schon alles kann und was ihm oder ihr zu einem komplexeren Ziel hin zu tun bleibt. Die Lehrperson bespricht dies in regelmässigen Abständen mit der Schülerin resp. dem Schüler, erfährt dabei Näheres über den individuellen Lernprozess und berät ihn oder sie in Bezug auf seine spezifische Situation. Indem Schülerinnen und Schüler selber eine Lernweg-Leit-Kartei erstellen, können sie ihre erledigten Aufgaben selber kontrollieren und den Lernweg sowie den Lernprozess reflektieren. Die Kartei wird von Zeit zu Zeit der Lehrperson zur Verfügung gestellt. Sie ergänzt das pädagogische Tagebuch und geht in die Beurteilung des individuellen Lernstandes des Schülers resp. der Schülerin ein.
- Partner- und Gruppenbewertung (z.B. Aufgabenkontrolle des Partners bei Partnerdiktaten, beim Vorlesen eines Textes, beim Abhören von Vokabeln usw.).
- Schülermitbeurteilung bei Klassenarbeiten und nicht-standardisierten Tests.

Kleber (ebd., p. 280f.) erwähnt, dass die Selbstbeurteilung durch Schülerinnen und Schüler günstige Bedingungen dafür bietet, dass die Anforderungen für die Betroffenen durchschaubarer werden, dass sie sich besser auf die Lern- und Prüfsituation einstellen können und ihnen durch klare Vorstellungen über die Leistungsmassstäbe die festgelegten Anforderungen eher erreichbar erscheinen. Neben der Leistungssteigerung, der Erhöhung der Lernbereitschaft sowie der kognitiven Kontrolle und Handlungskompetenz wirkt Selbstbeurteilung somit auch angst- und stressmindernd.

Bei der Selbstbeurteilung treffen, so Kleber (ebd., p. 279), pädagogische und diagnostische Forderungen hart aufeinander. Auf der pädagogischen Seite sind Mitbestimmung und Mündigkeit sehr bedeutend. Selbstkritik und realistisches Einschätzen der eigenen Möglichkeiten sollen geübt werden. Andererseits stellt sich aus diagnostischer Sicht die Frage, ob neben der ohnehin schon unzureichenden Objektivität und Zuverlässigkeit auch noch subjektive Betroffenheit und fehlende Selbstdistanz bei den Schülerinnen und Schülern als zusätzliche Fehlerquellen in Kauf genommen werden sollen. Man muss noch diskutieren, inwieweit diese Fehlerquellen vielleicht kontrolliert oder sogar vernachlässigt werden können.

Wild & Krapp (2001, p. 540f.) beschreiben, warum Selbstbeurteilung im schulischen Kontext nur selten anzutreffen ist: Obwohl durch Selbstbeurteilung die Mündigkeit des Lerners und die Fähigkeit zu eigenverantwortlichem Handeln als wichtige pädagogische Ziele geübt werden können, erschweren verschiedene Gründe diese Art der Beurteilung. So z.B. die möglicherweise nicht hinreichend entwickelte Fähigkeit der Schülerinnen und Schüler, die eigenen Leistungen sachgerecht einschätzen zu können oder die mangelnde Bereitschaft, offen schlechte Leistungsresultate oder Lernschwierigkeiten zu berichten. Ausserdem ist es für Schülerinnen und Schüler oft sehr schwer, eine richtige «Diagnose» zu erstellen, da nur eine in der Sache kompetente Person erkennen kann, ob das Wissen richtig und vollständig ist. Lernende sind erst ab einem gewissen Alter in der Lage, ein angemessen realistisches Bild der eigenen Fähigkeiten aufzubauen, eine völlig autonome Selbstbeurteilung ist also in den meisten Fällen gar nicht möglich. Es braucht die Unterstützung der Lehrperson. Keine Schwierigkeiten dürfte die Korrektur von einzelnen Arbeitsergebnissen (z.B. gegenseitiges Korrigieren eines Diktates) für die Schülerinnen und Schüler darstellen. Hier kann die stärkere Einbeziehung der Lernenden zum Teil erhebliche kognitive und motivationale Effekte erzielen.

In einer Studie von Lissmann und Paetzold (1984) wurden Auswirkungen verstärkter Selbstbeurteilung auf motivationale Faktoren, z.B. kompetitive vs. kooperative Einstellungen, getestet. In der Hauptschule (Schule mit geringsten Anforderungen) zeigte sich keine besondere Wirkung. Realschülerinnen und Realschüler hingegen zeigten durch vermehrte Selbstbeurteilung weniger Konkurrenzdenken und Schulangst. Das Selbstwertgefühl stieg und Misserfolge wurden häufiger mit von

der Schülerin und vom Schüler selbst beeinflussbaren Ursachen in Verbindung gebracht (vgl. Wild & Krapp 2001, p. 541).

Da durch die Selbstbeurteilung die Anforderungen für die Lernenden durchschaubarer werden und sie sich besser auf die Prüfsituation einstellen können, Prüfungsstress also vermindert wird, könnte die Selbstbeurteilung eine gute Ergänzung zu den herkömmlichen Beurteilungsformen sein. Sie bietet die Möglichkeit, die Schülerinnen und Schüler Mündigkeit und die Fähigkeit zu eigenverantwortlichem Handeln üben zu lassen. Es ist aber für die Lernenden sehr schwierig, eigene Leistungen sachgerecht zu beurteilen. Sie brauchen dazu die Hilfe der Lehrperson. Eine völlig autonome Selbstbeurteilung ist kaum möglich.

2.4.4 Portfolio

Portfolios kommen in Bereichen vor, in denen Leistungen nicht nur oder nur sehr schwer mit Ziffernnoten ausgedrückt werden können, so z.B. bei Künstlern, deren Fähigkeiten am ehesten mit Fotografien, Skizzen, Zeichnungen und Entwürfen dokumentiert werden können (vgl. Richter 2003, p. 44).

Im Schulbereich wurden Portfolios erstmals in den 80er Jahren in den USA in Writing Classes der Universitäten eingesetzt. Die Studierenden legten in Portfolios die Ergebnisse ihrer Arbeit eines Semesters ab. Ein Portfolio enthielt aber nicht nur die fertigen Essays, sondern auch erste Entwürfe, Ideen, Notizen und Überlegungen sowie (missglückte) Schreibversuche. Ein nicht unwesentlicher Teil der Arbeit bestand darin, zu überlegen, was die Schülerinnen und Schüler über ihr eigenes Schreibverhalten gelernt und wie der Unterricht ihren Schreibprozess unterstützt oder behindert hatte. Mittlerweile hielten Portfolios in den USA als alternatives Bewertungssystem in verschiedenen Fächern Einzug (ebd., p. 44).

Paulson, Paulson und Meyer (1991) definieren ein Portfolio wie folgt:

> «Ein Portfolio ist eine zielgerichtete Sammlung von Schülerarbeiten [z.B. gelöste Aufgaben, eine Arbeitsprobe oder ein Werk], welche die Anstrengungen des Lernenden, den Lernfortschritt und die Leistungsresultate auf einem oder mehreren Gebieten zeigt» (zit. nach Lissmann 2001, p. 487).

Gemäss Lissmann (2001) kann man sich ein Portfolio als einen «Behälter» zur Aufbewahrung verschiedener Arbeiten vorstellen, das kann z.B. ein Ordner, eine Schachtel oder eine CD-ROM sein. Die Schülerinnen und Schüler beteiligen sich bei der Auswahl der Inhalte und der Aufstellung der Kriterien für die Beurteilung.

Hacker (2003, p. 11) beschreibt, dass Portfolios anhand von ausgewählten Dokumenten Einblicke in das Können, den Arbeitsstil und die Entwicklung eines Autors resp. einer Autorin geben können. Ziel eines Portfolios ist es, nicht nur die Lernergebnisse sichtbar zu machen, sondern auch den eigenen *Lernprozess*. Bräuer (2003) nennt eine wichtige Funktion des Portfolios als alternative Bewertungsform: «Bewertet wird die Entwicklung [von] Fähigkeiten und Fertigkeiten über einen längeren Zeitraum hinweg betrachtet, nicht die Verfügbarkeit von Kenntnissen im Moment der Leistungsüberprüfung» (Bräuer, zit. nach Richter 2003, p. 46). Portfolios bieten den Lernenden einen Einblick in ihren individuellen Lernprozess, weil sie gezwungen sind, diesen zu reflektieren und zu kommentieren (vgl. Richter 2003, p. 45). Somit wird die Evaluation der eigenen Leistung gefördert. Diese reflexive Auseinandersetzung mit dem eigenen Lernen führt zu mehr Eigenverantwortung und Selbststeuerung im Lernen (vgl. ebd., p. 45).

Zu den Gütekriterien: Gemäss der klassischen Testtheorie kann eine Beurteilungsmethode, die nicht objektiv ist, auch nicht reliabel und valide sein (vgl. Abbildung 1.4). Bei der Portfoliobeurteilung ist laut Lissmann (2001, p. 491f.) Objektivität allerdings nur schwer realisierbar. Hacker bemerkt demgegenüber, dass sich die Güte eines Portfolios daran bemisst, wie konsequent die Prinzipien

Partizipation, Kommunikation und Transparenz realisiert werden (vgl. Hacker 2003, p. 10). Während man die festen Teile eines Portfolios relativ einfach beurteilen kann, ist es natürlich schwierig, einen Bereich wie «Grad der Selbstreflexivität» in Noten auszudrücken (vgl. Richter 2003, p. 46). Richter betont, dass nicht alle Probleme der Leistungsmessung und –bewertung durch Portfolios gelöst werden. So bedeuten Portfolios zunächst einmal erhebliche Mehrarbeit für alle Betroffenen. Allein die Planungszeit für die Lehrperson ist erheblich, da die Beurteilungskriterien offener gelegt werden als im herkömmlichen Unterricht. Auch die Schülerinnen und Schüler haben mehr Arbeit ausserhalb des Unterrichts, sei es nur bezüglich der angemessenen Präsentation der Ergebnisse. Allerdings bedeuten alle Neuerungen zu Beginn Mehrarbeit, bis eine gewisse Routine aufkommt. Schwerer wiegt das Problem bei mangelnder Selbständigkeit der Schülerinnen und Schüler. Es kann vielen schwer fallen, selbständig zu arbeiten und die eigene Leistung auch noch zu beurteilen (vgl. ebd., p. 46).

Die Portfoliobeurteilung eröffnet neue Möglichkeiten für Diagnose, individuelles Lernen und Schule. Es lassen sich Unzulänglichkeiten der traditionellen Beurteilung vermeiden, dies betrifft vor allem die Authentizität der Aufgaben, die Differenziertheit der Rückmeldungen, die Längsschnittbetrachtung der Leistung und die Multidimensionalität der Messung. Mit Hilfe der Portfoliobeurteilung können neue Leistungsanforderungen wie selbstgesteuertes Lernen oder selbstverantwortliches Lernen besser eingelöst werden (vgl. Lissmann 2001, p. 492 ff.). Demnach könnte also auch das Portfolio eine wichtige Ergänzung zu den herkömmlichen Beurteilungsformen sein.

Ein Beispiel für ein Portfolio – welches mittlerweile in vielen Schulen der Sekundarstufe II eingesetzt wird – stellt das Europäische Sprachenportfolio[44] dar. Die Idee stammt vom Europarat. Dieser hat dazu einheitliche Regeln formuliert. Das Schweizer Sprachenportfolio ist das erste Sprachenportfolio, das vom Europarat genehmigt wurde. Inzwischen gibt es mehrere Dutzend Sprachenportfolios in zahlreichen europäischen Ländern. Das Europäische Sprachenportfolio soll ein persönlicher Begleiter beim Sprachenlernen sein. Es soll die Mobilität in Europa erleichtern, indem es sprachliche Qualifikationen transparent und vergleichbar dokumentiert. Es besteht aus drei Teilen, dem Sprachenpass, der Sprachbiografie und dem Dossier. Der Sprachenpass ist das Kernstück des Sprachenportfolios und gibt einen Überblick über den aktuellen Stand der Sprachkenntnisse. Die Sprachbiografie dokumentiert die Geschichte des Sprachenlernens, hier werden interkulturelle Erfahrungen, Erlebnisse beim Benutzen der einzelnen Sprachen usw. notiert. Im Dossier werden schliesslich persönliche Arbeiten wie Texte, Briefe oder Referate gesammelt. Das Sprachenportfolio ist einerseits ein Referenzdokument, das es erlaubt, die Sprachkenntnisse vorzuzeigen, andererseits ist es ein Lernbegleiter, der zum Sprachenlernen motivieren und es autonomer werden lassen soll (vgl. EDK 2003b).

2.5 Notengebung in der Schule

Im Rahmen der schulischen Leistungsbeurteilung nehmen Noten eine besondere Stellung ein. Die Lehrperson ist in ihrer Funktion als Bewerterin und Beurteilerin dazu angehalten, erbrachte Leistungen zu bewerten und rückzumelden. Dies geschieht meist in Form von Zensuren bzw. Noten.[45] Es reicht nicht aus, festzustellen, wie viele Fehler eine Schülerin, ein Schüler gemacht hat, sondern es muss auch entschieden werden, ob die Leistung eine sehr gute, gute, befriedigende, ausreichende, mangelhafte oder ungenügende ist und dies gegenebenfalls in Noten zu übersetzen. Der folgen-

---

[44] Vgl. www.sprachenportfolio.ch

[45] Hinter dem Wort Zensur steckt zensieren – ein Begriff, der seinen Ursprung im lateinischen Verb censere (den Wert schätzen, begutachten) hat. In diesem Sinne ist eine Zensur zwar nicht zwangsläufig eine Ziffernzensur, sie stellt aber ausdrücklich eine Bewertung dar (vgl. Jachmann 2003, p. 23).

de Abschnitt befasst sich mit der Kritik an dieser Notengebung und stellt ausgewählte Ergebnisse der Forschung vor. Anschliessend wird in Abschnitt 2.5.3 geklärt, wieso die Notengebung trotz anhaltender Kritik immer noch das wesentliche Instrument zur Beschreibung der Leistungen darstellt.

2.5.1 Kritik an der Notengebung

Mit dem Titel «Die Fragwürdigkeit der Zensurengebung» veröffentlichte Ingenkamp 1971 einen äusserst erfolgreichen Sammelband. Er erreichte 1976 bereits die sechste und 1995 die neunte Auflage. Die in diesem Werk und der darauffolgenden Forschungsliteratur in Frage gestellte Notengebung ist einerseits eine Kritik an der Note als häufigste Form der Leistungsrückmeldung, betrifft aber andererseits auch den ganzen Prozess der schulischen Leistungsbeurteilung. Aus messstatistischen Gründen steht die Note als abhängige Variable[46] im Mittelpunkt der Forschungen und der Kritik. Aus der Vielfalt der Befunde werden im Folgenden zwei wichtige Dimensionen dargestellt: Die Subjektivität der Notengebung (Abschnitt 2.5.1.1) einerseits, das klassen- und fachinterne Bezugssystem (Abschnitt 2.5.1.2) andererseits.

2.5.1.1 Subjektive Fehlerquellen

Häufiger Vorwurf an die schulische Notengebung ist der, dass die Gütemassstäbe Objektivität, Reliabilität und Validität nicht erfüllt würden (vgl. Abschnitt 1.4). Lehrpersonen beurteilen nicht objektiv genug, d.h., das Messergebnis ist nicht vom Untersucher unabhängig. Folgende Forschungsergebnisse illustrieren diesen allgemein akzeptierten Befund[47]:

- In einer Untersuchung hat Ulshöfer (1963) von rund 1600 deutschen Abituraufsätzen zu einem zentral gestellten Thema die Benotung durch zwei unabhängige Lehrpersonen untersucht. Er fand, dass bei 12% der Aufsätze die beiden Urteile um mehr als eine Notenstufe differierten. Bei 21 Arbeiten lagen die Urteile sogar um mehr als zwei Notenstufen auseinander, derselbe Aufsatz wurde in diesem Fall also von einer Lehrperson mit einer zwei, vom anderen mit einer fünf bewertet (vgl. Konradt 1975, p. 118). Diese *mangelnde Auswertungsobjektivität* wurde wiederholt bestätigt, so auch bei Schröter (1971), der die Untersuchungsanlage auf über 11'000 Zensuren vergrösserte (vgl. Ingenkamp 1977, p. 148).
- *Implizite Persönlichkeits- und Charakterurteile* spielen eine gewisse Rolle, d.h., die Lehrperson lässt sich von Erwartungen, die sie an eine Arbeit heranträgt, lenken. Weiss (1965) liess Lehrpersonen Klassenarbeiten in Deutsch und Mathematik beurteilen, zusätzlich lieferte er einige fiktive Hintergrundinformationen über die zu beurteilenden Schülerinnen und Schüler.[48] Damit konnte gezeigt werden, dass bei 92 Lehrpersonen die Häufigkeit der zweitbesten Note für Rechtschreibung von 40% auf 7% der Urteile verschlechterte, wenn bei derselben Arbeit statt der positiven die negative Information zugrundegelegt worden war. Dies galt auch im umgekehrten Fall, wo sich die Häufigkeit, mit der die beste Note vergeben wurde, von 0% auf 11% erhöhte (vgl. ebd.).

---

[46] In einem statistischen Modell ist die abhängige Variable diejenige Variable, deren Ausprägungen durch eine oder mehrere andere Variablen – die unabhängige(n) Variable(n) – erklärt oder vorhergesagt werden sollen (vgl. ILMES 2006).

[47] Bei der Interpretation der Forschungsergebnisse ist jeweils zu berücksichtigen, dass es sich mehrheitlich um ältere Arbeiten aus Deutschland handelt. Auffälligster Unterschied zu den schweizerischen Beurteilungsformen ist, dass in Deutschland die Noten 1-4 für sehr gute, gute, befriedigende und ausreichende Leistungen, die Noten 5 und 6 für mangelhafte Leistungen eingesetzt werden (vgl. Sacher 2001, p. 47).

[48] «Diese besagten einmal, dass die Arbeit von einem relativ guten Schüler aus gutem Hause, zum anderen, dass sie von einem mittelmässigen, eher vernachlässigten Schüler stammt» (Konradt 1975, p. 119).

- Weniger bekannt ist, dass auch Lehrpersonen zu verschiedenen Zeitpunkten die gleiche Arbeit unterschiedlich bewerten. Ellis kam bereits 1930 zu dem Ergebnis, dass bei einer nach 11 Wochen wiederholten Bewertung der Antworten auf Geografie- und Geschichtsfragen durch 61 Lehrpersonen die meisten von ihnen kaum die Hälfte der Antworten gleich beurteilten. Ein Drittel der Lehrpersonen beurteilte sogar alle Fragen beim zweiten Mal anders (vgl. Konradt 1975, p. 119).

Eine Reihe weiterer subjektiver Fehlerquellen sind bis heute entdeckt worden. Als Effekte werden Inferenzfehler (Voreingenommenheiten) und Urteilsfehler genannt:

- *Halo-Effekt*: «Der allgemeine Eindruck einerseits, positive oder negative Beobachtungen oder Kenntnisse über den Schüler andererseits beeinflusst die Wahrnehmung von schulischen Fachleistungen» (Rhyn 2002; Dubs, Delhees & Metzger 1973).
- *Logische Fehler*: «Nicht selten zieht man voreilige Schlussfolgerungen von einem Leistungsmerkmal, das man schon kennt, auf ein anderes, erst zu beurteilendes» (Sacher 2001, p. 37). Dazu gehört z.B., wenn Lehrpersonen bei der Benotung eines Aufsatzes aufgrund vieler Rechtschreibfehler auch den Inhalt ungerechtfertigterweise schlecht bewerten.
- *Reihungseffekt*: «In einer Serie von schlechten Prüfungen werden die Arbeiten zunehmend besser benotet und umgekehrt» (Rhyn 2002, p. 26). Reihungsfehler ergeben sich aus dem Zusammenhang mit vorangehenden Urteilen. Dazu gehört auch der
- *Kontrasteffekt*: «Nach einer Serie von sehr guten Leistungen wird eine mittelmässige Leistung tendenziell als schlecht bewertet und umgekehrt» (ebd., p. 26).
- *Erste Arbeit*: «Die erste Arbeit einer Korrekturserie wird tendenziell anders bewertet als die letzte, wobei der Effekt mit der Länge der Gesamtkorrektur zunimmt» (Oelkers 2002, p. 5).
- *Tendenz zur Beharrlichkeit*: «Lehrpersonen rücken von einem einmal gefällten Urteil bei späteren Beurteilungen kaum ab» (Rhyn 2002, p. 26).
- *Strengefehler*: Strengefehler liegen vor, wenn bereits kleinere Mängel relativ stark gewichtet und vorwiegend schlechte Noten vergeben werden. Hier wird wie bei den anderen Urteilsfehlern das Beurteilungsspektrum nicht ausgeschöpft (vgl. Sacher 2001, p. 35).
- *Mildefehler*: Der Mildefehler zeichnet sich durch vorwiegend gute Noten aus.
- *Tendenz zur Mitte*: Bei der Tendenz zur Mitte werden Extremwerte vermieden, die Noten bewegen sich um 4.5 (vgl. Rhyn 2002, p. 26).
- *Wissen-um-die-Folgen-Fehler*: «Bei absehbaren negativen Folgen für die Schülerin oder den Schüler wird milder beurteilt» (ebd., p. 26). Lehrpersonen beurteilen möglicherweise auch milder, um Konflikte mit Eltern zu vermeiden.
- Wenn persönliche Beziehungen die Beurteilung beeinflussen, spricht man von *Fehlern durch Sympathie bzw. Antipathie* (vgl. ebd., p. 26).

2.5.1.2 Das klassen- und fachinterne Bezugssystem

Noten werden meist aufgrund des klasseninternen Bezugssystems gegeben. In einer leistungsstarken Klasse erhalten Schülerinnen und Schüler für die gleiche Leistung tiefere Noten als in leistungsschwachen Klassen. Ingenkamp (1995) hat dies z.B. für das Fach Mathematik nachgewiesen (vgl. Abbildung 2.4). Die Stabdiagramme geben an, welche Note die Schülerinnen und Schüler einer Klasse in einem standardisierten Rechentest für eine bestimmte Punktzahl erhielten (Note 2 = gut, oberes Stabende, Note 3 = befriedigend, mittlere Marke, Note 4 = ausreichend, unteres Stabende).

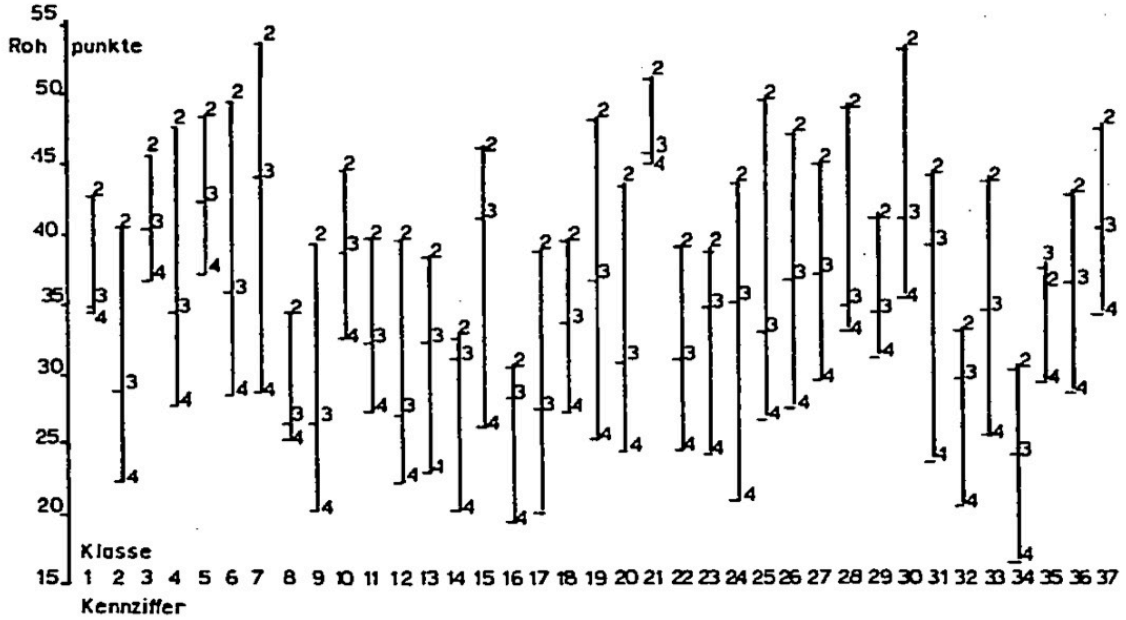

Abbildung 2.4: Problematik klasseninterner Bezugssysteme bei der Benotung (aus: Ingenkamp 1995, p. 197)

Eine Schülerin, die in der Klasse 21 eine 4 erhält, würde für die selbe Leistung in Klasse 32 mindestens eine 2 bekommen. Wenn sich ihre individuelle Leistung steigert, zugleich aber auch die Mehrheit der Klasse bessere Ergebnisse erbringt, so wird sie weiterhin unbefriedigende Noten erhalten. Als Konsequenz sind die Noten von Schülerinnen und Schülern aus verschiedenen Klassen selbst dann nicht mehr vergleichbar, wenn man die Möglichkeit einer objektiven Beurteilung innerhalb einer Klasse unterstellt.

Neben dem klasseninternen Bezugssystem ist auch ein fachinternes Bezugssystem wirksam; die Fächer werden verschieden streng bewertet: Hopp und Lienert (1965) haben in einer grösseren Untersuchung die Notenverteilung für die einzelnen Fächer im Gymnasium ermittelt. Sie fanden erhebliche Unterschiede in den Verteilungen. So entfiel die Note vier in Latein auf 40% der Fälle, in Biologie dagegen nur auf 11% (vgl. Konradt 1975, p. 119).

Wenn unter diesen Umständen Notendurchschnitte über alle Zensur-Fächer hinweg berechnet werden, so können die Leistungen eines Schülers mit einer eins in Latein und einer drei in Geografie nicht mit den Leistungen einer Schülerin mit einer drei in Latein und einer eins in Geografie verglichen werden. Erschwerend kommt hinzu, dass bei einer anderen Klasse die Notenverteilung anders sein könnte. Die Bewertungen und die für Selektionsverfahren (Aufnahme, Numerus Clausus etc.) wichtigen Noten variieren von Schule zu Schule, von Klasse zu Klasse (die Klassen 34 und 35 in Abbildung 2.4 sind aus demselben Schulhaus) und von Fach zu Fach zum Teil erheblich.

2.5.1.3 Fazit

Ergänzt man die Tatbestände der Kritik an der Notengebung mit weiteren, hier nicht aufgeführten Forschungsergebnissen, so muss man sich fragen, inwiefern eine gerechte(re) Notengebung überhaupt realisierbar ist. Die Bilanz ist ernüchternd:

- Die Beurteilungen der schulischen Leistungen zeichnen sich vielfach durch mangelnde Objektivität aus. Verschiedene Lehrpersonen beurteilen dieselbe Schülerinnen- oder Schülerleistung unterschiedlich; dies gilt für alle Unterrichtsfächer. Noten sind zwischen Lehrpersonen deshalb kaum vergleichbar.

- Eine Vielzahl subjektiver und leistungsirrelevanter Fehlerquellen wie z.B. die Tendenzen zu Strenge resp. Milde oder die Tendenz zur Mitte wirken sich auf die Note aus.
- Vorurteile und Erwartungen können die Beurteilung beeinflussen.
- In verschiedenen Fächern wird mit unterschiedlicher Strenge zensiert. «Die Strenge nimmt mit der Bedeutung des Faches zu, Hauptfächer haben die strengste Notengebung. Die Strenge der Benotung nimmt ebenfalls mit der Klassenstufe zu, die Gewichtung der Fächer und die entsprechende Strenge der Beurteilung bleiben aber gleich. In mild bewerteten Fächern werden oft nur die obersten Ziffern der Skala genutzt» (Rhyn & Moser 2002, p. 27f.).
- Bei der unterschiedlichen Benotung von verschiedenen Fächern spielt nicht nur der Rang des Faches eine Rolle. «In der Tendenz wird umso strenger zensiert, je mehr die Leistungen in schriftlichen Arbeiten überprüft werden, je mehr die Leistungen quantifizierbar sind» (ebd., p. 28).
- Mädchen erhalten tendenziell bessere Zeugnisnoten als Jungen. Sie repetieren weniger und haben inzwischen den grösseren Schulerfolg als Knaben (vgl. ebd., p. 28).

Aus all den oben angeführten Gründen ist es streng genommen nicht zulässig, aus einzelnen Noten Durchschnitte zu berechnen und diese miteinander zu vergleichen (vgl. Abschnitt 1.2.2).

Der nächste Abschnitt wendet sich nun der Frage zu, wieso trotz wiederholt nachgewiesener Nachteile von Prüfungen und Noten, trotz jahrzehntelanger Reformprojekte und trotz globaler Notenkritik seitens vieler Erziehungswissenschaftlerinnen und -wissenschaftler sowie Schulreformerinnen und -reformer das Notenschema praktisch unverändert eingesetzt wird.

2.5.2 Verbesserungsmöglichkeiten der schulischen Leistungsbeurteilung

An Hand der obigen Ausführungen wird klar, dass durch die Minimierung von Inferenz- und Beurteilungsfehlern ein grosser Beitrag zur Verbesserung der Beurteilung geleistet werden kann. Weitere Möglichkeiten liegen bei den folgenden Aspekten: die Art und das Zustandekommen der Beurteilung transparent und verständlich kommunizieren, die Vergleichbarkeit der Noten erhöhen oder überhaupt gewährleisten, sowie ein professioneller Umgang mit dem Dilemma der Förderung und Selektion erreichen. Wie wir in Abschnitt 2.5.3 sehen werden, machen die dort aufgeführten Vorteile der schulischen Notengebung klar, dass Vorschläge zur Verbesserung der schulischen Leistungsbeurteilung nur dann realistisch sind und mit breiter Akzeptanz rechnen können, wenn sie den Lehrpersonen keinen erheblichen Mehraufwand bescheren. Das heisst: Reformprojekte müssen auch die knappen zeitlichen Ressourcen der Lehrpersonen kalkulieren, lange Anforderungskataloge und unrealistische Forderungen sind zu vermeiden.

2.5.2.1 Inferenz- und Urteilsfehler vermindern

Inferenz- und Urteilsfehler können vermieden und die Objektivität der Notengebung erhöht werden, wenn die Lehrpersonen einige Massnahmen zur Vermeidung von Inferenz- und Urteilsfehlern berücksichtigen. Folgende Massnahmen können nach Sacher (2001) dazu beitragen, Inferenzfehler auszuschalten:

- Leistungsfeststellung und Leistungsbeschreibung sind von der Leistungsbewertung möglichst zu trennen.
- Ein Gesamturteil soll immer erst gebildet werden, nachdem genügend Daten zu Einzelaspekten gesammelt wurden. Z.B. ist bei einer Klassenarbeit zunächst die erste Aufgabe über die ganze Klasse hinweg zu korrigieren, erst dann folgt die Korrektur der zweiten Aufgabe.
- Ganzheitliche Leistungen wie Aufsätze, Produkte des Kunstunterrichts und Übungen an Turngeräten sind besonders anfällig für Inferenzfehler. Kriterien können dazu dienen, die Leistung in Teilaspekte zu unterteilen.

- Vorgängig ist zu überlegen, welche Aspekte der Leistung beurteilt werden sollen, die Beurteilung soll nicht von Zufälligkeiten gelenkt werden.
- Die Reihenfolge bei Korrekturarbeiten ist zu variieren, um so die Reihungsfehler zu umgehen.
- Das Bild, das man von einer Schülerin oder einem Schüler hat, muss regelmässig überprüft und das Urteil gegebenenfalls revidiert werden (vgl. Sacher 2001, p. 37f.).

Um Urteilsfehlern zu begegnen, sind folgende Grundsätze zu beherzigen:

- Eine Lehrperson sollte sich Klarheit darüber verschaffen, zu welchem Fehlverhalten sie neigt. Konkret kann die Häufigkeit der gegebenen Urteile notiert, eine Klassenarbeit durch mehrere Lehrpersonen beurteilt und das Gespräch mit Kolleginnen und Kollegen gesucht werden.
- Die alterstypische Leistungsfähigkeit korrekt einschätzen lernen: Dabei hilft die entwicklungspsychologische Fachliteratur, Gespräche mit Kolleginnen und Kollegen und der Einsatz standardisierter Schulleistungstests.
- Grundsatz sollte sein, dass in jedem Leistungsbild nach Stärken und Schwächen gesucht wird (vgl. Sacher 2001, p. 35f.).

Das Wissen um die üblichen Fehlerquellen wie den Halo-Effekt, die Tendenz zur Beharrlichkeit, den Reihungs- und den Kontrasteffekt, die Tendenz zur Mitte, der Wissen-um-die-Folgen-Fehler und Fehler durch Sympathie bzw. Antipathie sowie das Wissen um die grundsätzliche Begrenztheit des Lehrerinnen- und Lehrerurteils sind wichtige kognitive Grundlagen für Lehrpersonen. Wenn daraus eine verantwortungsbewusste Einstellung im Umgang mit Beurteilungen und Noten erzeugt werden kann, so wäre dies eine Verbesserung der Beurteilungspraxis (vgl. Mattl 1978, p. 122). Aber: Die idealen Beurteilerinnen und Beurteiler resp. Notengeberinnen und -geber wird es nie geben. Lehrpersonen werden weiterhin die Leistungen der Schülerinnen und Schüler verzerrt wahrnehmen. Das Prinzip «je genauer, desto besser» ist daher zu relativieren.

2.5.2.2 Transparente Leistungsbeurteilung und Notengebung

Ein wichtiger Ansatz zur Verbesserung der schulischen Leistungsbeurteilung liegt darin, eine transparentere schulische Leistungsbeurteilung einzurichten. Unterrichts- und Beurteilungsformen sollten allen Beteiligten klar offengelegt und transparent kommuniziert werden. Es soll verständlich und nachvollziehbar sein, welche Stärken und Schwächen die einzelnen Unterrichts- bzw. Beurteilungsformen enthalten, welchen Aufgaben und Anliegen sie gerecht werden können und welchen nicht.

Eine transparente Leistungsbeurteilung und Notengebung zeichnet sich durch folgende Merkmale aus:

- Entwicklung transparenter Lernprogramme mit möglichst realistischen Zeiterwartungen. Die Schülerinnen und Schüler können dann individuell den eigenen Lernaufwand, von dem der Lernerfolg massgeblich abhängig ist, abklären und einteilen (vgl. Oelkers 2002, p. 14).
- Minimallernziele, Kriterienkatalog, Bezugsnorm und Bewertungsmassstab werden offengelegt.

2.5.2.3 Einsatz standardisierter überregionaler Schulleistungstests

Regional durchgeführte Vergleichsarbeiten[49] sind für Lehrpersonen wichtig, wenn eine Standortbestimmung vorgenommen werden soll. Die Lehrperson kann sich so Klarheit über den Leistungs-

[49] Helmke & Hosenfeld (2004, p. 2) definieren Vergleichsarbeiten wie folgt: «Es handelt sich um *schriftliche Arbeiten*, die in einer grösseren Anzahl von Schulen (ggfs. landesweit) auf der Basis einer vorgegebenen Aufgabenstichprobe eingesetzt werden mit dem Ziel, die Leistungen der Schüler an einer klassen- und schulübergreifenden sozialen und/oder kriterialen Bezugsnorm zu messen. [...] Grundsätzlich [wird] die eine Hälfte der Aufgaben der Vergleichsarbeiten zentral (vom Ministerium) festgelegt, während die andere Hälfte

stand ihrer Klasse(n) verschaffen. Aufgrund des Vergleiches der Ergebnisse mit dem «real gemessenen Leistungsstand» der Schülerinnen und Schüler ist zu beurteilen, ob es sich um eine im regionalen Vergleich eher leistungsschwache, durchschnittliche oder leistungsstarke Klasse handelt. Um die Aussagekraft der Bewertung vergleichbarer und transparenter zu machen, bietet sich folgendes Verfahren an: «Drei Lehrpersonen unterrichten das gleiche Fach […]. Sie stellen gemeinsam Prüfungen zusammen, lassen diese in den drei Klassen schreiben, korrigieren und bewerten die Ergebnisse nach abgesprochenen Regeln und Massstäben» (Vögeli-Mantovani 2001, p. 21). Eine solche Standortbestimmung ist jedoch noch keine Gewährleistung dafür, dass die Beurteilungen nach aussen hin vergleichbarer werden. Dies ist erst dann gegeben, wenn Tests zumindest überregional durchgeführt werden. So wie z.B. das in Abschnitt 2.6.3 vorgestellte Klassencockpit.

Seit mehr als 30 Jahren wiederholt Ingenkamp seine Forderung (vgl. Ingenkamp 1971), die traditionelle Beurteilungspraxis (Klassenarbeiten, mündliche Prüfungen und Beobachtungen) und das für die Beschreibung der Leistungen vorherrschende Instrument der Notengebung durch ein besseres diagnostisches Verfahren, durch standardisierte Tests, zu ersetzen oder zu ergänzen. Solche Schulleistungstests gibt es in Deutschland mittlerweile für alle Schulstufen, in der Schweiz bestehen weiterhin nur wenige standardisierte überregionale Schulleistungstests.

Folgende Elemente können als Indizien dafür gelten, dass sich dies ändern könnte:

- Die Teilnahme der Schweiz an den auf standardisierten Wissenstests basierenden internationalen Schulleistungsvergleichsstudien TIMSS und PISA.
- Finanzierungsprobleme des Staates zwingen zu Sparmassnahmen, die eingesetzten Mittel sind vermehrt zu legitimieren. Bezüglich der Bildung heisst dies: «Überprüfung der Effektivität (Wirksamkeit) der öffentlichen (staatlich finanzierten) Schulen, Nachweis der Qualität der Bildungseinrichtungen und Steigerung der Effizienz der staatlichen Mittel» (Herzog et al. 2004, p. 4).
- Mittel sollen nur noch eingesetzt werden, wenn die Effektivität gesteigert wird und sich der Aufwand lohnt. Der Wandel von der bisherigen Input- zu einer solchen Output-Steuerung impliziert Evaluation und Qualitätssicherung. Die interne Evaluation oder Selbstevaluation der Lehrpersonen soll mit externer Evaluation durch unabhängige Instanzen ergänzt werden und damit eine vergleichende Beurteilung von Schulen oder Schulsystemen ermöglichen. Das Strategieziel der EDK, ein nationales Bildungsmonitoring aufzubauen (vgl. EDK 2001, p. 2), ist in diesem Wandel der politischen Steuerungskultur zu sehen.
- Im Juni 2002 stimmte die EDK dem Projekt «Harmonisierung der obligatorischen Schule HarmoS» zu. Das Projekt hat einen wichtigen Bezug zum EDK-Arbeitsschwerpunkt «Aufbau eines nationalen Bildungsmonitorings» und verpflichtet diejenigen Kantone, die dem entsprechenden Konkordat beigetreten sind, Ziele und Strukturen der obligatorischen Schule anzugleichen (vgl. EDK 2003a). Das Konkordat ist seit 2009 in Kraft, 15 Kantone sind bisher beigetreten.
- Die Ergebnisse einer Befragung von 206 Lehrerinnen und Lehrern über ihre Erfahrungen im Rahmen der Erprobung von Klassencockpit (vgl. Abschnitt 2.6.3) im Schuljahr 2002/03 sind sehr positiv (vgl. Moser 2003). Die Mehrheit begrüsst das Klassencockpit und wünscht sich ein Instrument zur Standortbestimmung und zur Verbesserung der Unterrichtsqualität (vgl. ebd., p. 15). Die Hauptintention von Klassencockpit liegt bei der Standortbestimmung. Die Lehrpersonen haben die Ergebnisse jedoch zu verschiedenen weiteren Zwecken wie «Bestätigung des

---

von den Grundschulen aus dem vorgegebenen Aufgabenpool ausgewählt wird. Dabei folgt die Auswahl einem Schlüssel, der gewährleistet, dass unterschiedliche Lehrplanbereiche und Kompetenzklassen über alle Schulen hinweg in vergleichbarer Weise repräsentiert sind.» Sie grenzen Vergleichsarbeiten von folgenden sechs Instrumenten zur schulischen Leistungsbeurteilung ab: Parallelarbeiten, Orientierungsarbeiten, standardisierte Schulleistungstests, Lernstandserhebungen, Klassenarbeiten und Klassencockpit (vgl. ebd.).

Leistungstandes der Klasse, Orientierung der Eltern und des Kollegiums, Nutzung der Informationen zur Verbesserung des Unterrichts [und zur] Anpassung der Schülerbeurteilung» (Moser 2003, p. 16) genutzt.

- In eine ähnliche Richtung weist der von Neuenschwander, Hirt, Balmer, Gasser, und Goltz (2003) in einem Forschungsprojekt entwickelte standardisierte Schulleistungstest[50] für die Klassenstufen 6 und 8. Der Test ist frei verfügbar; dies soll es den Lehrerinnen und Lehrern des Kantons Bern ermöglichen, die Tests in ihren Klassen einzusetzen und die «Leistungen ihrer Schülerinnen und Schüler in Beziehung zu setzen zu den Leistungen der Schülerinnen und Schüler jener Klassen, die an der Untersuchung beteiligt waren» (ebd., p. 1). Das Einsatzziel des Tests sehen die Autorinnen und Autoren darin, dass er eine Standortbestimmung der Leistung der eigenen Klasse ermöglicht.

Die Klassencockpit-Evaluation hat gezeigt, dass Lehrpersonen einen professionellen, lehrplankompatiblen standardisierten Leistungstest begrüssen. Der Nutzen scheint klar: Klassencockpit orientiert über den Leistungsstand der Klasse und bietet sich so als Instrument zur Standortbestimmung, zur Objektivierung der Beurteilung und sogar zur Optimierung des Unterrichts an. Unter diesen Voraussetzungen wäre es wünschenswert, wenn Lehrpersonen mehr standardisierte «Schulleistungstests als bisher zur Verfügung stehen, die sie dann, wenn sie es für sinnvoll halten und wenn es den behandelten Sachgebieten entspricht, einsetzen können» (Mattl 1978). Die Idee der überregionalen Leistungsmessungen soll nicht als ein übertriebener, bürokratischer Aufwand gesehen werden, der nur zu Statistikzwecken erfolgen soll, sondern als ein Versuch, eine wirklich brauchbare und gerechte Bezugsnorm zu erreichen, auf die sich alle Lehrpersonen beziehen können. Die Bezugsnorm wäre damit überregional dieselbe und in diesem Sinne objektiv, zuverlässig, unbeeinflussbar und auch allgemein gültig. In diesem Sinne kann sie auch im Kleinen, also vor Ort für die einzelne Schule, für die einzelne Lehrperson, brauchbar sein und Anwendung finden. Dadurch können die durch Noten suggerierte Objektivität, Reliabilität und Validität sowie die klassenübergreifende Vergleichbarkeit der Leistungs-Beurteilung erhöht werden.

2.5.2.4 Förderorientierte und selektionsorientierte Notengebung

Im Folgenden wird aufgezeigt, dass die Entflechtung der sich dilemmatisch gegenüberstehenden Aufgaben der Schule, nämlich Förderung und Selektion, ein wesentlicher Beitrag zu Transparenz und Verbesserung der schulischen Leistungsbeurteilung sein kann.

Selektionsorientierte und förderorientierte Leistungsbeurteilung lassen sich als Gegensatzpaare darstellen und bilden in diesem Sinne eine Dichotomie, die es erlaubt, entgegengesetzte Standpunkte einzunehmen und zu vertreten. Dies lässt sich vor allem anhand der reformpädagogischen Rhetorik aufzeigen, wo oft ein Wandel von selektionsorientierter hin zu förderorientierter Beurteilung gefordert wird. Dieser Wandel wird hauptsächlich mit der Veränderung des Unterrichts erklärt. Dabei soll die Tatsache, dass sich immer mehr individualisierende Unterrichtsformen in den Schulen etabliert haben, dazu führen, dass herkömmliche Beurteilungssysteme nicht mehr genügen können.[51] Die selektionsorientierte Leistungsbeurteilung und damit vor allem die Beurteilung mit Ziffernoten wird als die bestehende und, trotz allen Reform- und Veränderungsbestrebungen, immer noch als

---

[50] «Der Test wurde nach Kriterien der klassischen Testtheorie zur summativen Überprüfung von Leistungen in Mathematik entwickelt» (Neuenschwander et al. 2003, p. 33).

[51] Plössnig (1997) konstatiert – wie andere Autorinnen und Autoren auch – eine Veränderung der Unterrichtspraxis in Richtung differenzierter und individualisierender Unterricht. Zeugnisnoten vermögen demgegenüber nicht mehr zu bestehen und deshalb ist eine Alternative erforderlich. «Zeugnisnoten, wie sie allgeallgemein üblich sind, eignen sich aus mehreren Gründen nicht für eine neue, offene Unterrichtsgestaltung, in der auf jedes einzelne Kind Rücksicht genommen wird, wo jedes Kind best-mögliche Förderung erhalten sollte» (Plössnig 1997, p. 663f.).

die geltende Beurteilungsform kommuniziert. Die förderorientierte Beurteilung wird indes als *die* Alternative, gleichsam als gegenüberliegender oder gegensätzlicher Pol dargestellt, der im Rahmen selektionsorientierter Beurteilung scheinbar keine oder zu wenig Beachtung findet.

Interessant ist auch, wie mit der Möglichkeit umgegangen wird, dass beide Beurteilungsformen ihre Berechtigung haben und in diesem Sinne wichtige Aufgaben der Schule darstellen. Einige Autorinnen und Autoren plädieren für einen Ersatz der selektionsorientierten Leistungsbeurteilung und damit vor allem der Notengebung durch die förderorientierte Leistungsbeurteilung. Andere wiederum sehen die Notwendigkeit, dass weder das eine noch das andere vernachlässigt werden sollte und sprechen sich für eine Ergänzung aus. Die Debatte scheint also von der Kontroverse beherrscht zu werden, ob beide Formen nebeneinander bestehen können, oder ob sie einander ausschliessen.

Die Diskussion über «Ersatz» oder «Ergänzung» läuft sich tot und kann zur Lösung des Problems nur sehr wenig beitragen. Es ist deshalb für die Offenlegung des Problems hilfreich und entscheidend festzuhalten, dass zwar auf rein argumentativer und rhetorischer Ebene das eine gegen das andere ausgespielt werden kann, wenn es darum geht, für einen Ersatz zu plädieren, dass man aber auf der anderen Seite vor einem Dilemma steht, wenn es darum geht, beide Anliegen im Unterricht zu vereinen. Genau dies ist aber die Aufgabe eines öffentlichen Bildungssystems.

Die Problematik, welche sich aus den beiden Beurteilungsformen ergibt, stellt sich nur deshalb, weil Unterricht nicht von Leistungsbeurteilung getrennt werden kann. Für die Schule beinhaltet eine handlungsbezogene Diagnostik sowohl den Unterricht und seine Formen, wie auch die darauf abgestimmten Arten der Leistungsmessung und der Leistungsrückmeldung. Unterricht und Beurteilung müssen in diesem Sinne kohärent und aufeinander abgestimmt sein. Wenn sich der Unterricht hin zu einer individualisierenden und förderorientierten Form ändert, muss sich demzufolge zwangsläufig auch die Art der Beurteilung ändern. Wenn wir auf der anderen Seite aber davon ausgehen, dass auch selektionsorientierte Beurteilung von der Schule – wie oben angedeutet – wahrgenommen werden *muss*, steht die Schule vor dem Problem, für zwei Beurteilungsformen auch zwei Unterrichtsformen bieten zu müssen. Sie muss zwangsläufig jedem Unterricht eine bestimmte und darauf abgestimmte Beurteilungsform zuordnen oder jeder Beurteilungsform einen bestimmten Unterricht vorausschicken. Krapp (1994) formuliert dies folgendermassen:

> «Für unseren Zusammenhang ist die Feststellung wichtig, dass Modifikations- und Selektionsstrategien wegen ihrer unterschiedlichen Zielsetzungen, Vorannahmen und Lösungsmuster jeweils andere Anforderungen an die handlungsbezogene Diagnostik stellen. Es müssen deshalb auch jeweils andere Theorien und Methoden zur Bewältigung dieser diagnostischen Aufgaben herangezogen werden» (Krapp 1994, p. 574).

Konkret wird damit die Problematik der beiden Dimensionen Fördern und Auslesen angesprochen. Für die Selektion bedeutet dies, dass der Unterricht so angelegt sein muss, dass Leistungsbeurteilungen vergleichbar sind. Eine förderorientierte Beurteilung andererseits ist unmöglich, wenn im Unterricht nicht individualisiert und in einem gewissen Sinne «ungleich» be- oder eben gehandelt wird. Schule muss sowohl Chancengleichheit und Beurteilung nach dem Leistungsprinzip als auch Berücksichtigung individueller Unterschiede und Förderung individueller Fähigkeiten gewährleisten.

Die Paradoxie entsteht erst dann, wenn klar wird, dass von der Schule nicht nur bestimmte Beurteilungsformen, sondern auch Unterrichtsformen verlangt werden. In diesem Sinne darf sie nicht bloss eine innere Reform vollziehen und auf eine veränderte Klientel eingehen, indem sie den Unterricht individualisiert. Sie muss auch Funktionen und Aufgaben übernehmen, welche von aussen, also strukturell bedingt, an sie herangetragen werden. Eine solche systemisch bedingte Anforderung stellt die Selektion dar, weil sie anlässlich eines vertikal gegliederten Oberstufensystems nicht einfach übergangen werden kann.

Vögeli-Mantovani (1998) bietet eine mögliche Lösung des Problems an. Seiner Meinung nach sollen die beiden Anliegen der Individualisierung bzw. der Förderung und der Selektion entflochten werden.[52] Während des Schuljahres soll demnach förderorientiert oder lernprozessunterstützend beurteilt werden. Die Selektion hingegen soll erst am Ende des Schuljahres, zu einem bestimmten Zeitpunkt, stattfinden.[53] Lernprozessunterstützende Massnahmen können ohne einen individualisierenden Unterricht schlecht getroffen werden. Die förderorientierte Beurteilung kann damit also nach einem darauf abgestimmten Unterricht erfolgen, da er individualisierend stattfinden kann, und erfüllt in diesem Sinne die Forderung nach Kohärenz zwischen Unterrichts- und Beurteilungsform. Es stellt sich nun aber die Frage, ob es zulässig ist, nachdem der Unterricht individualisierend stattgefunden hat, eine Selektion und damit eine kollektive Beurteilungsform durchzuführen. Die Schülerinnen und Schüler haben im Rahmen förderorientierten Unterrichtes keineswegs dieselben Bedingungen erfahren. Nur wenn alle an der Selektion Beteiligten denselben selektionsrelevanten Unterrichtsstoff vermittelt bekommen haben, können auch kollektive Schultests zur Anwendung kommen. Dabei ist nicht entscheidend, dass der Unterricht für alle genau gleich ablaufen muss, sondern dass alle den gesamten, für einen erfolgreichen Abschluss relevanten, Unterrichtsstoff bearbeitet haben. Den Unterricht zu individualisieren bedeutet aber, für jedes Individuum andere, namentlich den Lernprozess fördernde, Unterrichtsbedingungen schaffen zu müssen. Das würde aber auch heissen, dass nicht alle alles machen müssten und dass daher nicht alle genau gleich gut auf einen Test vorbereitet würden. Dies wiederum würde aber bedeuten, dass bereits im Voraus eine verborgene oder versteckte Selektion stattgefunden hat. Die Selektion setzt also ihrerseits auch bestimmte Anforderungen an den vorangegangenen Unterricht. Die Kohärenz zwischen Unterrichts- und Beurteilungsform ist also mit einer Entflechtung nicht in beiden Fällen gewährleistet. Die Frage, ob ein förderorientierter und individualisierender Unterricht mit einer selektionsorientierten Beurteilung vereint werden kann oder nicht, bleibt folglich trotz dieses plausiblen und logisch erscheinenden Vorschlags weiterhin offen und unbeantwortet.

Das Dilemma bleibt somit bestehen. Die schulische Leistungsbeurteilung kann sich weder auf Förderung beschränken, noch darf sie sich ausschliesslich am Selektionsauftrag orientieren. Dies zeigt auch das folgende Zitat:

> «Die eigentliche Schwierigkeit der Notengebung liegt also darin, dass Noten nach aussen hin vergleichbar und objektiv sein sollen, um für alle Schüler, die sich um Ausbildungs- und Studienplätze bemühen, die gleichen Voraussetzungen zu schaffen, dass sie aber für interne, schulische Belange eigentlich flexibel sein müssten. Dabei heisst flexibel nicht, dass sie willkürlich anwendbar sein sollten, vielmehr ist damit gemeint, dass Lehrer Bewertungsschlüssel auf ihren Unterricht hin ausrichten können […], dass sie auf die individuelle Situation der Schüler Rücksicht nehmen können» (Mattl 1978, p. 122).

2.5.2.5 Fazit und Ausblick

Die oben aufgeführten Verbesserungsmöglichkeiten sind in keiner Weise umfassend und streifen die aktuelle Diskussion nur kursorisch. Es stellt sich die Frage, inwiefern die von erziehungswissen-

---

[52] «Damit Schülerbeurteilung ihren Zweck erfüllt, müssen die verschiedenen Formen und Funktionen klar auseinander gehalten werden. Insbesondere ist die formative von der summativen und prognostischen Beurteilung zu unterscheiden» (Vögeli-Mantovani 1998, p. 32).

[53] «Die Beurteilung des Lernens und der Entwicklung in Unterricht und Schule soll im Schuljahr vorherrschend sein und umgehend dem Lernen und der Entwicklung zugute kommen. Die Promotion von einer Klasse zur nächsten und die Selektion in die Sekundarstufe I und II sind zu bestimmten Terminen auf eine umfassende und prognostische Beurteilung angewiesen. Eine Entflechtung meint, dass im Schuljahr vorwiegend die formative, lernprozessunterstützende Beurteilung zum Zuge kommen soll und dass eine summative, ausleseorientierte Beurteilung auf die vorgegebenen Termine beschränkt werden soll» (Vögeli-Mantovani 1998, p. 31).

schaftlicher Seite gestellten hohen Anforderungen an die Schülerinnen- und Schülerbeurteilung nicht zu einer Überforderung der Lehrpersonen führen. In diesem Zusammenhang wirkt die folgende Aussage versöhnlich:

> «Es gibt für die Unterrichtsarbeit im Klassenzimmer keine didaktischen Modelle, keine speziellen Lehrmethoden und keine rationalen Aufgabenzuweisungen an die Schüler, die durch extreme Genauigkeit der herangezogenen diagnostischen Informationen wesentlich verbessert werden könnte. Das gleiche gilt natürlich in verstärktem Masse für die häufig geforderte Übereinstimmung der Lehrerurteile mit vergleichbaren Testwerten. Wichtig allein ist eine ungefähre Diagnose der Lehrerurteile und ihre permanente Überprüfung im Verlauf des Unterrichts» (Helmke 2003, p. 89).

Es ist nicht zwingend notwendig, dass die Lehrperson möglichst präzise diagnostische Urteile über den je aktuellen Leistungsstand bilden und dies so mitteilen kann. Unter Umständen ist es pädagogisch gar günstiger, wenn sich die Lehrperson an der individuellen Bezugsnorm orientiert und kleine Fortschritte auch auf tiefem Niveau honoriert, wenn sie regelmässig hohe Ansprüche stellt, und sich trotz anhaltenden Misserfolgen durch eine optimistische Erfolgshaltung auszeichnet und nicht resigniert. Es spricht einiges dafür, dass nicht die exakte Einschätzung, sondern eine leicht positiv verzerrte Überschätzung des Leistungspotentials der Schülerinnen und Schüler positiv auf die effektive Leistung auswirkt. Dies führt mit grösserer Wahrscheinlichkeit dazu, dass den Schülerinnen und Schülern anspruchsvolle, gerade noch im Rahmen ihrer Bewältigungsmöglichkeiten liegende Aufgaben gestellt und ihnen so günstige Entwicklungsanreize geboten werden (vgl. Schrader 2001, p. 92). Diagnosen von Lehrpersonen während des Unterrichts müssen, im Gegensatz zu landläufigen Überzeugungen, keineswegs besonders genau sein, wenn sich der Diagnostiker der Ungenauigkeit, Vorläufigkeit und Revisionsbedürftigkeit seiner Urteile bewusst ist (vgl. Helmke 2003, p. 89).

2.5.3 Rehabilitierung der Notengebung

Die empirische Notenforschung verfolgt Fragestellungen, die seit Beginn des 20. Jahrhunderts stabil sind. Voten gegen Prüfungen und Noten finden sich schon am Ende des 19. Jahrhunderts, seinerzeit vorgebracht von Medizinern, die die Fachöffentlichkeit mit den schädlichen Folgen von Prüfungsstress konfrontierten. Zwischen 1950 und 1970 lässt sich dann ein starkes Forschungsaufkommen nachweisen, das danach im deutschen Sprachraum im Unterschied zu den angelsächsischen Verhältnissen abnimmt (vgl. Oelkers 2002, p. 2f.). Ingenkamp (1971) fasste in seinem Sammelband diese Forschungsergebnisse zusammen. Verschiedene neuere Studien befassen sich mit den von ihm beschriebenen Dimensionen der «subjektiven Fehlerquellen», dem «klasseninternen Bezugssystem», den «Qualitätsunterschieden in den einzelnen Schulfächern» und dem «prognostischen Wert» der Notengebung. Der Topos der «Fragwürdigkeit» und die damit verbundene grundsätzliche Kritik an der Beurteilungspraxis und der Notengebung blieben aber unverändert (vgl. Oelkers 2002, p. 4).

Dieser fundamentalen Kritik an der Notengebung muss aber entgegengehalten werden, dass das Objektivitätsideal nie hinterfragt und Gegenevidenzen nie erzeugt worden sind. Auffällig ist auch, dass die grundlegenden Datensätze kaum erneuert wurden. Die Forschung berücksichtigt also in ungenügendem Masse, ob und wie sich die Beurteilung und die Notengebung verändert hat. So ist es wenig plausibel, von einer seit 30 Jahren unveränderten Sprach- und Aufsatzdidaktik auszugehen. In den 90er Jahren waren Zeugnisse mit Noten in fast allen Kantonen noch die Regel (vgl. Vögeli-Mantovani 1999a, 1999b). Inzwischen machten aber alle Kantone von ihrem Recht, eigene Wege in der Beurteilungspraxis zu suchen, zu finden und zu gehen, ausgiebig Gebrauch. Diese Reformschübe führten zu einem für das föderalistische Bildungssystem der Schweiz üblichen vielfältigen Bild (vgl. ebd.). Die Schule hat die Notwendigkeit von verbesserten Formen der Leistungsbeurteilung eingesehen. Inzwischen gibt es nicht nur Noten, sondern auch Alternativen (vgl. Oelkers 2002, p. 15):

- Noten mit und ohne explizite Bezugsnormen;
- Lernberichte;
- fakultative wie nicht-fakultative Beurteilungsgespräche;
- Orientierungsarbeiten zur Standortbestimmung;
- Selbstbeurteilungen der Schülerinnen und Schüler;
- Zeugnisse mit lernzielbezogenen Wortetiketten;
- Zeugnisse mit lernzielbezogenen Wortetiketten für Beurteilung des Lernprozesses und der Leistung.

Es ist wissenschaftshistorisch interessant, dass die Ergebnisse der 60er und 70er Jahre weiterhin unverändert herangezogen werden, wenn es darum geht, die Notengebung zu kritisieren. Zumindest die Fragestellungen müssten dem aktuellen Stand der Beurteilungspraxis angepasst werden.

Dies genügt jedoch nicht: Auch nach jahrzehntelanger Kritik ist das *hauptsächliche* Instrument zur Beschreibung von Leistungen immer noch das Notenschema: Obwohl Noten umstritten sind, haben sie für die Beurteilung der Schülerinnen und Schüler kaum an Bedeutung verloren. Dieser Befund ist so deutlich, dass die grundlegende Frage gestellt werden muss, warum es trotz klarer Missstände immer noch Prüfungen und Noten gibt. Die Praxis der Notengebung muss für die Schule und die Lehrpersonen eine Vielzahl von Vorteilen haben, dass sie trotz der von Ingenkamp initiierten und von der Schulreformdiskussion z.T. aufgenommenen anhaltenden Kritik weiterhin und hartnäckig dabei bleiben.

Der Wert der Note liegt besonders in seiner *zweckdienlichen* Eigenschaft als Berechtigungs- und Auslesefaktor (vgl. Jachmann 2003, p. 43). Noten sind *kurz, eindeutig und einfach zu entziffern*, da sie, im Gegensatz zu Wortgutachten, keiner weiteren alltagssprachlichen Interpretation bedürfen. Die Probleme der Ausdeutung sind gering. Noten verursachen des Weiteren keine *grösseren Probleme bei der Formulierung negativer Beurteilungen*. Die Alternative «Textliche Leistungsbeschreibung» oder «Wortgutachten» verlangen von Lehrpersonen einen erheblich grösseren Aufwand. Mit Noten *«kann man» Durchschnitte berechnen*, Einzelnoten lassen sich zu *Gesamtnoten* zusammenfassen. Noten gelten in Selektionsverfahren als *relativ schnell und handlich*, denn sie *erlauben einfache Vergleiche* und *ermöglichen die Bildung von Leistungshierarchien*. Die Kritik, dass die Notenskala nur eine Ordinal- bzw. Rangskala ist (vgl. Abschnitt 1.2.2) und deshalb keine Mittelwerte berechnet werden dürfen, würde entfallen, wenn die Gleichheit der Skalenabstände (Äquidistanz) vorhanden wäre.

Der Vorteil der Noten ergibt sich auch aus den Nachteilen der zahlreichen, als Alternativen formulierten Beurteilungsverfahren wie Wortgutachten, Selbstbeurteilung und Einschätzskalen. Alternativen sind vorhanden, aber in allen Varianten tauchen wiederum notenartige Schematisierungen auf: «Im Prinzip bleiben die Schwierigkeiten, die mit der Benotung verbunden sind, bestehen, wenn verbale Beurteilungen vorgenommen werden. Zahlen werden durch Worte ersetzt. Dadurch verändert sich die Kommunikation der Beurteilung, nicht aber die Beurteilung selbst» (ebd., p. 33). Eine weitere Schwierigkeit von Wortgutachten besteht darin, dass die durch Lehrpersonen vorgenommenen alltagssprachlichen Einschätzungen häufig der Interpretation bedürfen. Es besteht die Gefahr von Uneindeutigkeit. Auch der Aufwand, der betrieben werden muss, um Wortgutachten zu erstellen, ist vergleichsweise gross. Noten hingegen sind einfache Instrumente, sie lassen sich auf *ökonomische Weise einsetzen und kommunizieren* (vgl. Oelkers 2002, p. 2).

Kritik am psychologischen Wert von Noten soll weiterhin diskutiert werden, der z.T. fragwürdige Einsatz der Noten als Motivierungs- und Disziplinierungsmittel sowie die mathematische Scheingenauigkeit dürfen und müssen breit diskutiert werden. Daraus folgende Reformprogramme oder Programme zur Erhöhung der Diagnosefähigkeit der Lehrpersonen müssen aber daraufhin getestet

werden, inwiefern sie nur den Aufwand der Lehrpersonen erhöhen, die Messgenauigkeit jedoch ihrerseits nicht zu steigern vermögen (vgl. ebd., p. 13).

Schülerinnen und Schüler sowie ihre Eltern wünschen sich aussagekräftige Beurteilungen – am liebsten mit Zensuren und Kommentaren gleichermassen. Mit zunehmendem Alter wünschen sich die Jugendlichen häufiger ein Zeugnis mit Noten (vgl. Rhyn & Moser 2002, p. 34). Eine Kombination aus Noten- und Berichtszeugnis könnte das Notenzeugnis mit separatem Kommentarbogen bieten. Auch von Schülerinnen und Schülern erstellte Portfolios wären als Ergänzung der Ziffernzeugnisse hilfreich.

2.6 Standardisierte Schulleistungstests

Standardisierte Schulleistungstests wurden in den siebziger Jahren auf Drängen der Pädagogik und der Öffentlichkeit unter die Lupe genommen (vgl. Ingenkamp 1997, p. 106). Ein zentrales Merkmal standardisierter Schulleistungstests ist, dass das Ergebnis einer geprüften Klasse mit den Ergebnissen einer bereits bestehenden repräsentativen Vergleichsgruppe in Beziehung gesetzt werden kann. Standardisierte Schulleistungstests können einerseits fächerübergreifend angelegt sein oder sich andererseits mit fachspezifischen Kompetenzen wie Rechtschreibung, Wortschatz, Rechnen oder Fremdsprachen befassen. Der Begriff standardisiert bezieht sich auf die Vereinheitlichung der Untersuchungsbedingungen, des Untersuchungsmaterials und der Auswertung (vgl. ebd., p. 118). Im Gegensatz zu den in der Schule geläufigen nicht-standardisierten Tests sind bei standardisierten Schulleistungstests auf Grund ihrer Konstruktion und der Standardisierung der Durchführung, Auswertung und Interpretation die Gütekriterien Objektivität, Validität und Reliabilität gewährleistet. Da sie sich auf die Curricula verschiedener Schulen und Regionen beziehen, sind sie inhaltlich generalistisch. Verfügbare Instrumente sind in einschlägigen Testkompendien (z.B. Borchert, Knopf-Jerchow & Dahabashi 1991), Testkatalogen (Hogrefe 1998) und Testdatenbanken ausführlich dokumentiert. Beispielsweise enthält die wichtigste deutschsprachige Testdatenbank «Psyndex»[54] Verweise auf 282 standardisierte Schulleistungstests (vgl. Wild & Krapp 2001, p. 538).

Die Konstruktion von standardisierten Schulleistungstests erfolgt anhand der Regeln der klassischen Testtheorie. Standardisierte Schulleistungstests lassen sich in zwei Hauptarten unterteilen (vgl. Ingenkamp 1997, p. 117):

A) Bezugsgruppenorientierte Schulleistungstests: Testverfahren, die individuelle Testergebnisse oder Testergebnisse einer Klasse mit den Resultaten einer repräsentativen Bezugsgruppe vergleichen.

B) Kriteriumsorientierte Schulleistungstests: Testverfahren, die angeben, ob die Lehrziele erreicht werden. Individuelle Testergebnisse oder Ergebnisse einer Klasse werden zu vorher gesetzten Kriterien in Beziehung gesetzt.

Im Folgenden werden diese beiden Formen standardisierter Schulleistungstests dargestellt. Abschliessen wird diesen Teil die Darstellung des Klassencockpits, ein standardisierter bezugsgruppenorientierter Schulleistungstest.

2.6.1 Standardisierte bezugsgruppenorientierte Schulleistungstests

Bei standardisierten bezugsgruppenorientierten Schulleistungstests[55] wird das individuelle Ergebnis einer Schülerin resp. eines Schülers (oder einer ganzen Schulklasse) mit den Resultaten einer relevanten Stichprobe verglichen und anhand von Prozentrangplätzen oder T-Werten (vgl. Abschnitt

[54] Vgl.www.zpid.de.

[55] Lukesch (1998) verwendet hierfür den Begriff (sozial-)normorientierter Schulleistungstest.

1.5 Normen) ausgewiesen (vgl. Ingenkamp 1997, p. 119). Lukesch (1998) definiert den standardisierten bezugsgruppenorientierten Schulleistungstest als ein

> «wissenschaftliches Routineverfahren zur Festlegung des Kenntnisstandes in einem (oder mehreren) inhaltlich spezifizierten kognitiven Lehrzielbereich(en); dabei werden Aussagen über die Leistungshöhe aufgrund des Vergleiches mit den Leistungen einer für die jeweilige Alterstufe, Schulstufe oder Schulart repräsentativen Stichprobe getroffen» (ebd., p. 501).

Die folgende Abbildung 2.3 zeigt schematisch den Weg von der Planung zur Auswertung eines standardisierten bezugsgruppenorientierten Schulleistungstests:

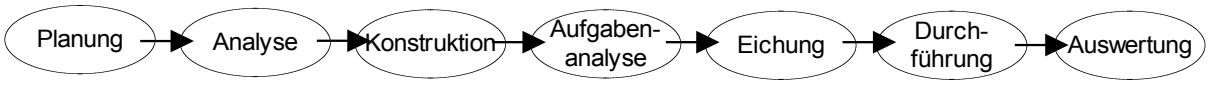

Abbildung 2.5: Entwicklungsphasen standardisierter bezugsgruppenorientierter Schulleistungstests (eigene Darstellung)

Es werden bestimmte Phasen von der Planung zur Auswertung unterschieden: In der *Planungsphase* wird geklärt, wer, was, warum misst. In der *Analysephase* filtert ein Expertenteam Lehrziele heraus und überträgt die wichtigsten in eine Matrix. Die Ziele werden bspw. anhand der Anforderungskriterien «Erinnern», «Verstehen» und «Denken» formuliert. Als nächstes folgt die *Konstruktionsphase*, in der Aufgaben für den Test entwickelt werden. Dabei entscheidet man, welcher Aufgabentyp den Zielsetzungen, Inhalten und Anwendungsmöglichkeiten am besten entspricht. Die nächste Phase ist die der *Aufgabenanalyse*. Dabei werden die Aufgaben einer kleinen Stichprobe (200-400 Schülerinnen und Schüler) vorgelegt, um etwa die Aufgabenschwierigkeit und die Trennschärfe zu berechnen. Diese Daten geben den Testkonstrukteuren eine Fülle von Informationen, anhand derer sie sich über die definitive Aufgabenauswahl entscheiden. Am Ende der Aufgabenanalyse wird die definitive Aufgabensammlung zusammengetragen. Diese werden einer repräsentativen Stichprobe vorgelegt, um nötige Vergleichsnormen zu erhalten (*Eichung*). Anschliessend wird der standardisierte Test zur Publikation oder zum Einsatz an einer Prüfung freigegeben. Der Test steht nun zur *Durchführung* und (computerunterstützten) *Auswertung* bereit.

Zentral für die Güte eines standardisierten bezugsnormorientierten Schulleistungstests ist die Analyse von Lehrplänen, speziellen Curricula und von Lehrbüchern für das jeweilige Fach. In diesem Zusammenhang wird von curricularer Validität gesprochen, die das Ausmass erfasst, «in dem Testitems ein vorgegebenes Curriculum repräsentieren» (Lukesch 1998, p. 514). Ebenso von Bedeutung ist aber die Lerngelegenheit, d.h. der Grad, in dem die Schülerinnen und Schüler *tatsächlich* Gelegenheit hatten, die in einem Schulleistungstest (und im Curriculum) repräsentierten Inhalte zu lernen. Sind beide Bedingungen gegeben, kann von einem inhaltlich validen Test gesprochen werden (vgl. ebd.).

Gemäss Lukesch (1998, p. 517) existieren folgende Einsatzmöglichkeiten des bezugsnormorientierten Tests:

- Vergleich des Leistungsstandes der ganzen Klasse mit der Eichstichprobe;
- Überprüfung des eigenen Benotungssystems durch den Vergleich mit den Testwertklassen resp. der geeichten repräsentativen Stichprobe;
- Objektivierungsmöglichkeit bei Schulartwechsel;
- Einsatz zur Lehr- und Lernsteuerung;
- Einsatz zur Unterrichtsdifferenzierung;
- Äussere Differenzierung (Einteilung in Leistungsgruppen);
- Summative Evaluation (Schulleistungstest als Ersatz für Klassenarbeiten).

2.6.2 Standardisierte kriteriumsorientierte Schulleistungstests

Beim standardisierten kriteriumsorientierten Schulleistungstest wird das individuelle Ergebnis in Relation zu einem vorher gesetzten Kriterium gesetzt. Abbildung 2.6 zeigt den Weg von der Planung zur Auswertung.

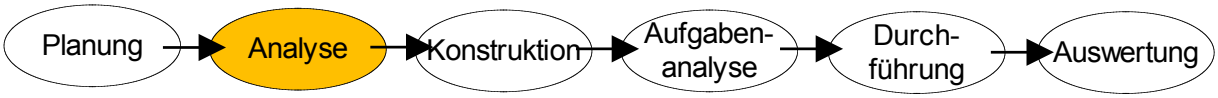

Abbildung 2.6: Phasen standardisierter kriteriumsorientierter Schulleistungstests (eigene Darstellung)

Die *Planung*, *Aufgabenkonstruktion*, *Aufgabenanalyse* und *Durchführung* unterscheiden sich formal kaum von der bezugsgruppenorientierten Testentstehung. Die *Eichung* entfällt, da keine Vergleichswerte an einer Bezugsgruppe gewonnen werden müssen. Der *curricularen Analyse* kommt hingegen besondere Bedeutung zu. «Das Problem besteht darin, das Universum all der Aufgaben, die die zu prüfende Unterrichtseinheit repräsentieren, genau zu erfassen und zu beschreiben und daraus eine repräsentative Aufgabenstichprobe für den Test zu ziehen» (Ingenkamp 1997, p. 126). Die Verfahren, die eine hohe Inhaltsgültigkeit garantieren sollen, sind dabei in der Einschätzung Ingenkamps (ebd., p. 125) wesentlich genauer als die Schätzungen von Expertinnen und Experten beim bezugsgruppenorientierten Schulleistungstest.[56] Bei der Bestimmung der Resultate (*Testauswertung*) müssen – im Unterschied zur bezugsgruppenorientierten Messung, wo die Daten anhand von Tabellen in Normen und Standardwerte umgerechnet werden – bei der kriteriumsorientierten Messung weitere Aspekte beachtet werden.

So muss bestimmt werden,

- bei welchem Wert das Kriterium erreicht wurde und
- in welche Leistungsgruppe, bezüglich des Abstandes vom Lernziel, der bzw. die Lernende einzuordnen ist.

Es gibt verschiedene Testauswertungssysteme (vgl. Ingenkamp 1997, p. 128), die mit Hilfe der Binomialverteilung (vgl. Bortz 1999, p. 65ff.) bzw. daraus abgeleiteten Tabellen angeben, wie viele Aufgaben eine Schülerin resp. ein Schüler mindestens lösen muss, damit man annehmen darf, dass er mit einer festgelegten Wahrscheinlichkeit alle zum Lernziel gehörenden Aufgaben lösen würde (vgl. Ingenkamp 1997, p. 128).

Standardisierte Schulleistungstests, sowohl bezugsgruppenorientierte als auch kriteriumsorientierte, sollen Schwächen der nicht-standardisierten Prüfungsmethoden vermeiden und werden meistens überregional eingesetzt. Als bekanntestes Beispiel für einen standardisierten Schulleistungstest gilt im Moment PISA.

2.6.3 Klassencockpit – Darstellung eines ausgewählten standardisierten Verfahrens

2.6.3.1 Hintergrund

Die Erziehungswissenschaft interessiert sich vermehrt für die «gute Schule» und fragt sich, welche *Qualität* diese ausweisen soll (vgl. Aurin 1991). Qualität ist ein Attribut der Marktwirtschaft, das auf die Schulen übertragen werden soll. Entsprechend geht es darum, Instrumente zu entwickeln,

[56] Zu erwähnen sind etwa das Verfahren der behavioristischen Lernzieloperationalisierung, die Tyler-Matrix, Lernzieltaxonomien etc. (vgl. Ingenkamp 1997, p. 126f.).

die als Hilfen in der Qualitätssicherung der Schulen eingesetzt werden können. Das St. Galler Projekt Klassencockpit ist eine Antwort auf diese Entwicklung. Es handelt sich um ein standardisiertes Testinstrument zur Orientierung über die Schülerleistung und zur Standortbestimmung für Lehrpersonen (vgl. Landert Farago Davatz & Partner 2002, p. 21). Das Klassencockpit wurde durch den Erziehungsrat des Kantons St. Gallen initiiert und über eine Projektgruppe konzipiert und umgesetzt. Im Kanton St. Gallen wurde das Klassencockpit bereits im Jahr 1995 das erste Mal bei einer kleinen Gruppe von Lehrpersonen eingesetzt, seither aber laufend weiterentwickelt. Die Projektgruppe wird wissenschaftlich durch das Kompetenzzentrum für Bildungsevaluation und Leistungsmessung der Universität Zürich betreut und in didaktisch-methodischen Fragen von Fachexperten der PH St. Gallen sowie von einer pädagogischen Kommission unterstützt. Die Modulentwicklung erfolgt durch eine Kerngruppe von Fachpersonen, die durch wechselnde Mitarbeitende eines Pools von Lehrpersonen unterstützt werden. Die betreffenden Lehrpersonen werden dafür vom Unterricht freigestellt. Mittlerweise wird das Klassencockpit in den Kantonen AI, AR, SH, TG, BS, ZH, NW, AG, GR sowie im Fürstentum Liechtenstein und in Privatschulen eingesetzt. Ziel war es, ein Instrument zu entwickeln, das Veränderungen im Bereich der schulischen Leistungsdiagnostik anstrebt und zur Qualitätssicherung in der Schule beiträgt. Dabei erhalten Lehrpersonen die Möglichkeit, durch den Einsatz dieses Instruments ihre diagnostische Kompetenz zu fördern und somit die Qualität des Unterrichts durch regelmässige Selbstevaluation zu sichern und weiterzuentwickeln. Die Lehrpersonen können selber über die Durchführung entscheiden und unterliegen keinerlei Auflagen der Schule oder der Schulkommissionen. Die Ergebnisse werden auch nicht zu Aufsichts- oder Kontrollzwecken benutzt, und auch die Veröffentlichung der Resultate ist verboten.

2.6.3.2 Anwendung

Das System beinhaltet momentan Aufgabensammlungen, so genannte Module, in den Fächern Deutsch und Mathematik für die 3. bis 9. Klasse. Die Inhalte der Leistungstests werden an den Gemeinsamkeiten der Lehrpläne der Deutschschweiz ausgerichtet (vgl. Muster 2003, p. 8ff.). Dreimal im Jahr haben die Lehrpersonen die Möglichkeit, ein Modul von Klassencockpit einzusetzen. Die Module können von den Lehrpersonen in Form eines Tests flexibel in das Unterrichtsprogramm integriert werden. Die Arbeiten der Schülerinnen und Schüler bleiben in der Hand der Lehrperson. Die Resultate werden mit den Ergebnissen einer repräsentativen Stichprobe von zur Zeit rund 500 Schülerinnen und Schülern verglichen.

2.6.3.3 Auswertung

Die Auswertung der Ergebnisse erfolgt über eine Online-Datenbank.[57] Nachdem die Lehrperson die Ergebnisse des Tests in eine Maske auf der Datenbank eingegeben hat (sie kann diese Daten auch auf einem Kontrollbogen ausfüllen), erhält sie sofort einen Vergleich der eigenen Klasse mit der Eichstichprobe. Die Auswertung führt zu einer Grafik, wie sie in Abbildung 2.7 ersichtlich ist.

[57] Vgl. www.klassencockpit.ch.

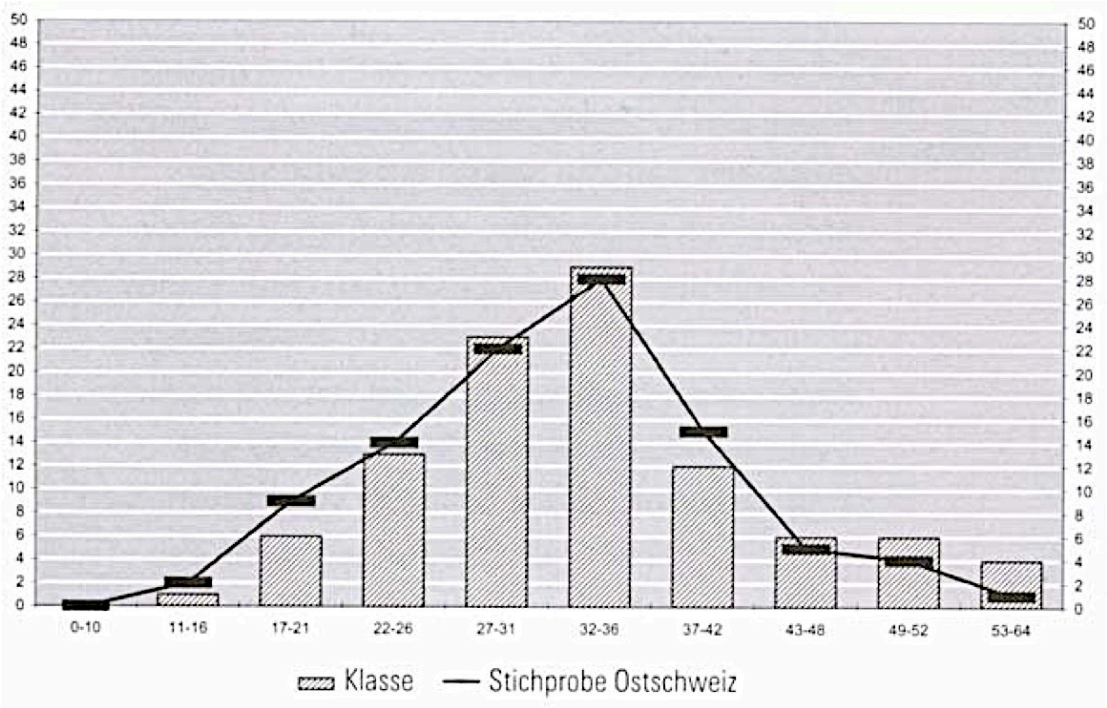

Abbildung 2.7: Grafik der Verteilung der Leistungen (aus: Zoppi 2003, p. 23)

Die Y-Achse zeigt die Anzahl der Lernenden in Prozent an, die X-Achse gibt Auskunft über die erreichte Punktzahl. Die Grafik zeigt die Verteilung der Leistungen innerhalb der Klasse (Balken), die mit der Verteilungskurve der Leistungen der Eichstichprobe (schwarze Linie) verglichen wird (vgl. Moser 2003, p. 10).

Neben der obigen Grafik liefert die Auswertung eine Tabelle, in der die Ergebnisse der Lernenden in Form der Punktzahl, des Prozentrangs und des Niveaus aufgelistet werden. Auch ein Notenvorschlag wird für die Beurteilung der Leistung angeboten. Tabelle 2.2 veranschaulicht diese Darstellungsform.

Tabelle 2.2: Bsp. eines Ausschnitts aus den Detailangaben für Lehrpersonen (aus: Zoppi 2003, p. 23)

Name	Punkte	Niveau	Prozentrang	Note
Schüler 1	9	3	20%	3.5
Schülerin 2	20	1	99%	5.5
Schülerin 3	11	2	35%	4
Schüler 4	23	1	99%	6

Die Aufgaben eines Moduls werden für drei Anforderungsniveaus entwickelt. Erreicht eine Schülerin oder ein Schüler Niveau 1, genügt sie/er erhöhten Anforderungen. In Niveau 2 erfüllen die Lernenden die Anforderungen teilweise, in Niveau 3 nur in geringem Masse. Die erreichte Punktzahl bestimmt das Niveau. Der Prozentrang besagt, wie viele Lernende der Eichstichprobe eine gleich gute oder schlechtere Leistung erbracht haben. Er informiert damit über die relative Stellung der Schülerinnen und Schüler in Bezug zur Eichstichprobe (bezugsgruppenorientierte Norm).

Durch das Klassencockpit können die Lehrpersonen somit verschiedene diagnostische Informationen erhalten (vgl. Zoppi 2003, p. 26f.):

- Orientierung über den Leistungsstand der Klasse (Verteilung, Leistungsspanne);
- Standortbestimmung (im Vergleich mit der Eichstichprobe);
- Interpretation von Ergebnissen einzelner Schüler und Schülerinnen;
- Kontrolle der eigenen diagnostischen Kompetenz (beim Vergleich eigener Ergebnisse mit Ergebnissen aus Klassencockpit).

Zudem wird das Klassencockpit von den Lehrpersonen auch als Grundlage für Elterngespräche und zur Hilfe bei der schulischen Selektion beigezogen (vgl. Moser 2003, p. 10). Generell wird die Auswertung der Ergebnisse im Internet von fast allen Lehrpersonen als verständlich, informativ und nützlich bezeichnet (vgl. ebd., p. 11f.). Die meisten Lehrpersonen schätzen ausserdem den geringen Aufwand, der mit der Auswertung verbunden ist. Auch die Durchführung der Tests bereitet den Lehrpersonen keine Mühe (vgl. ebd., p. 8).

2.6.3.4 Interpretation und Nutzung der Ergebnisse

Dem Prozess der Interpretation der Ergebnisse kommt eine wichtige Bedeutung zu (vgl. Kunz-Heim 2002, p. 27). Er beeinflusst die Qualität der Beurteilung erheblich. Um die Bedeutung der Ergebnisse genau bestimmen zu können, müssen die Messresultate von jeder Lehrperson in einen bestimmten Bezugsrahmen gesetzt werden. An diesem Punkt scheinen die Unsicherheiten und die Unzulänglichkeiten des Klassencockpits am grössten zu sein. So hält Moser (2003) fest:

> «Wenn Klassencockpit für eine Standortbestimmung eingesetzt wird, dann spielt der Zeitpunkt der Durchführung eine gewisse Rolle. Der Vergleich der Klasse mit der repräsentativen Stichprobe erfolgt mit Vorteil zum festgelegten Zeitpunkt. Nicht weniger von Bedeutung für die Ergebnisse in den Leistungstests ist, ob die geprüften Fachbereiche im Unterricht behandelt worden sind. Zudem spiegelt sich die Qualität des Unterrichts im Ergebnis. Eine weitere Erklärung der Ergebnisse liegt zweifelsohne bei den Lernvoraussetzungen der Schülerinnen und Schüler. Schulen unterscheiden sich in ihrem Einzugsgebiet in Bezug auf soziale und kulturelle Herkunft der Schülerinnen und Schüler, zwei Faktoren, die massgeblich für den Lernerfolg in der Schule verantwortlich sind. Dieser Sachverhalt wird in der Berechnung der Ergebnisse nicht berücksichtigt, er muss deshalb bei der Interpretation der Ergebnisse einer Klasse beachtet werden» (Moser 2003, p. 5).

Jede Lehrperson entscheidet selbst, wozu sie die Ergebnisse verwendet. Die Befragung von 216 Lehrpersonen des Kantons Zürich hat dabei ergeben, dass das Klassencockpit vor allem zur Bestätigung des Leistungsstandes der Klasse und zur Ergänzung der Schülerbeurteilung eingesetzt wird:

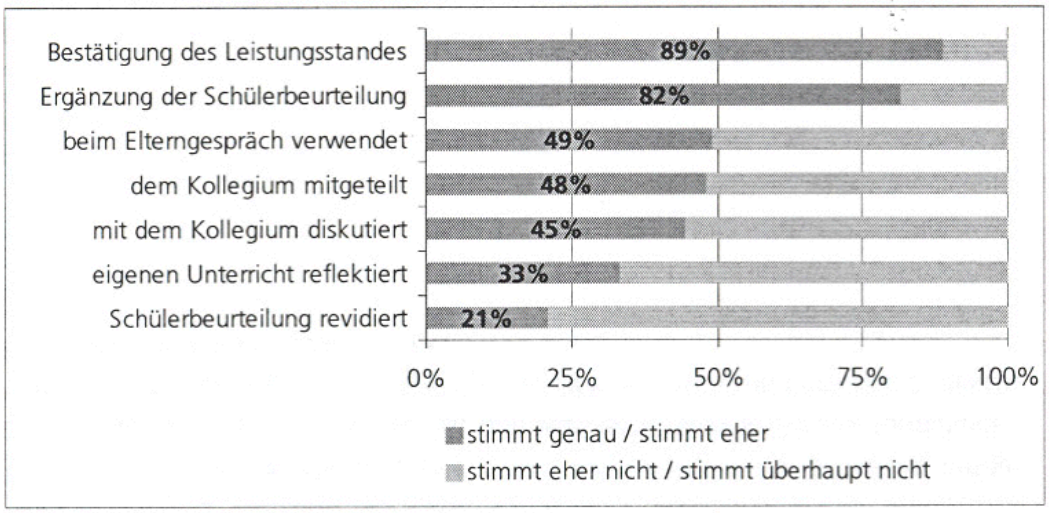

Abbildung 2.8: Nutzung der Ergebnisse von Klassencockpit (aus: Moser 2003, p. 14)

2.6.3.5 Gütekriterien zum Klassencockpit

Da das Klassencockpit nach wissenschaftlichen Kriterien entworfen worden ist, muss es den Kriterien der klassischen Testtheorie standhalten. Alle Gütekennzahlen wurden durch ein spezielles Expertenteam überprüft (vgl. Zoppi 2003, p. 31ff.):

Zur Objektivität:

- Durchführungsobjektivität: Durch die hohe Standardisierung der Tests und standardisierten Anweisungen wird Objektivität in der Durchführung gewährleistet. Sie ist dann eingeschränkt, wenn Lehrpersonen die Module den Schülerinnen und Schülern nach Hause mitgeben oder gezielt auf die Module hin üben.
- Auswertungsobjektivität: Die Aufgaben unterliegen eindeutigen Korrektur- und Bewertungskriterien. Standardisierte Handreichungen mit präzisen Lösungsschlüsseln führen zu einer hohen Objektivität bei der Auswertung. Die Auswertung ist weit gehend standardisiert.
- Interpretationsobjektivität: Sie wird dadurch verringert, dass jede Lehrperson die Ergebnisse selber interpretieren muss. Aus den gleichen Prüfungsergebnissen können unterschiedliche Schlüsse gezogen werden. Die Interpretationsobjektivität ist somit verringert.

Zur Validität:

- Inhaltsvalidität: Das Klassencockpit hätte dann eine hohe Inhaltsvalidität, wenn es die Fach- und Sachgebiete der verschiedenen Lehrpläne in repräsentativer Weise widerspiegeln würde. Fähigkeiten und Inhalte, die gemäss den Richtzielen der Lehrpläne geprüft werden sollen, werden deshalb von Fachleuten und Lehrpersonen bestimmt und in einem Konzept beschrieben. Nicht erfasst werden durch das Klassencockpit soziale, emotionale und komplexe kognitive Leistungsdimensionen.

Zur Reliabilität:

- Reliabilitätskoeffizienten: Die Trennschärfen der einzelnen Aufgaben sind grösser als 0.3 Punkte, was eine mittlere Höhe darstellt. Cronbachs Alpha, das den Bezug aller Aufgaben auf das zu messende Lehrkonstrukt misst, kann mit 0.8 als gut beziffert werden.
- Interraterreliabilität: Die Interraterreliabilität beschreibt, inwiefern zwei oder mehrere Beurteiler zum gleichen Ergebnis kommen. Dank dem Miteinbezug des Computers zur Übertragung der Daten in eine Datenmatrix und der computerisierten Auswertung ist die Interraterreliabilität kontrollierbarer und dadurch eher gegeben.

2.6.3.6 Fazit

Das Instrument Klassencockpit wird von den Lehrpersonen gut akzeptiert. Es dient der Orientierung über die Schülerleistungen und der Selbstevaluation des Unterrichts, resp. der diagnostischen Kompetenz. Die Beurteilung des Klassencockpits fällt in der Evaluation von Moser (2003) und Landert et al. (2002) positiv aus. Einschränkend muss jedoch gesagt werden, dass das Klassencockpit auch in nicht intendiertem Sinne eingesetzt wird: So üben einige Lehrpersonen mit ihren Klassen für die konkreten Aufgaben des Klassencockpits, das Klassencockpit wird zur Hausaufgabe umfunktioniert, es führt zu einem Schattennotensystem, die Deutung schlechter Ergebnisse wird bloss als Ergebnis einseitiger, scheinbar nicht lernzielkonformer Aufgaben betrachtet, die Schulbehörde zwingt zum Einsatz des Instruments etc. (vgl. Landert et al. 2002, p. 22). Hier muss zur Beseitigung dieser Fehlentwicklung die Kommunikation mit den betroffenen Personen und Behörden gesucht werden. Das Klassencockpit ist zwar nicht geeignet, sämtliche Bedürfnisse der schulischen Leistungsdiagnostik zu befriedigen, dennoch scheint es geeignet als begleitendes Diagnoseinstrument für Schule und Unterricht.

2.7 Basic-Check und Multi-Check: Leistungsbeurteilung durch ausserschulische Instanzen

In diesem Abschnitt wird eine neuere Tendenz im schweizerischen Bildungssystem thematisiert, wo Leistungsbeurteilung durch ausserschulische Instanzen zumindest ergänzt wird.

> «Selektion ist unabdingbar […]. Nimmt die Schule diese Funktion nicht mehr wahr, verschwindet nicht einfach die Selektion, sondern sie wird vom Arbeitsmarkt übernommen» (Criblez 2002, p. 87).

Diese Tendenz gilt vermehrt für die Besetzung von Lehrstellen. Den Ziffernzeugnissen wird – wie wir gesehen haben – zu Recht mangelhafte Vergleichbarkeit über die Klassengrenze hinaus angelastet. Dies scheint für grössere Firmen, die jährlich mehrere hundert Dossiers zu beurteilen haben, eine ungenügende Daten- bzw. Entscheidungsgrundlage zu sein. Bei Swisscom z.B. beginnen jährlich 30-40 Schulabgängerinnen und Schulabgänger ihre Lehre, die Auswahl an Bewerbungsdossiers liegt bei 500 (vgl. Gerber 2004, p. 14f.). Neben dem Zeugnis und dem Bewerbungsschreiben verlangt Swisscom noch die Ergebnisse eines Eignungstests wie den Basic-Check oder den Multicheck[58], bevor sie valable Kandidatinnen und Kandidaten zu einem Mini-Assessment einlädt. Die Lehrlingsausbildnerin der Swisscom betont, dass die zusätzlichen Testergebnisse das Bild ergänzen. Dennoch: wer bei einer national oder sogar international agierenden Firma eine Lehrstelle erhalten will, muss neben guten schulischen Noten auch ein entsprechendes Testergebnis vorweisen können.

Ausserschulische Instanzen, wie die einzelnen Lehrbetriebe, Berufsverbände oder aber unabhängige Firmen, versuchen die Lücke der mangelhaften Vergleichbarkeit von Zeugnisnoten zu schliessen und entwickeln sogenannte Eignungstests. Der Basic-Check wird von einer privaten Firma geführt. Als «Grundlagentest des Schulwissens für Lehrstellenbewerberinnen und Lehrstellenbewerber in Berufsbereichen wie Automation, Elektronik, Informatik, Konstruktion, Mechanik, Technik, Dienstleistung und Administration» (Amt für Berufsbildung und Berufsberatung des Kantons Solothurn 2006) soll der Basic-Check nebst Gesprächen, Noten und Schnupperlehren zu einem Bestandteil der Lehrlingsselektion werden. Durchgeführt wird der Test in einem der ca. 50 Abklärungszentren in der Schweiz oder in Liechtenstein; die Kosten von CHF 100.- sind durch die Schülerinnen und Schüler selbst zu tragen. Der Test wird am Computer durchgeführt und dauert ca. 3 Stunden. Die Multiple-Choice-Antworten werden zentral ausgewertet und die Jugendlichen erhalten ihre Resultate innert 5 bis 7 Tage per Post zugesandt (vgl. Erziehungsdirektion des Kantons Bern 2007; Basic-Check 2004). Die Schülerinnen und Schüler müssen sich zudem identifizieren. Dies erlaubt eine gewisse Kontrolle, denn der Test darf maximal zweimal durchgeführt werden. Im Jahr 2002 haben bereits 8000 Schülerinnen und Schüler den Basic-Check absolviert (vgl. Knellwolf & Vögeli-Mantovani 2003). Dies ist bei einer gesamtschweizerischen Anzahl von noch ca. 70'000 Lehrstellen jährlich (vgl. BBT 2004) eine nicht mehr zu unterschätzende Grössenordnung. Für die auf die Lehre vorbereitenden Schulen besteht die Gefahr, dass nicht mehr der Stoffplan, sondern die Inhalte des Basic-Check zum bestimmenden Unterrichtsgegenstand und damit zum wirksamen Lehrplan werden. Damit könnte sich eine informelle und vor allem von inoffizieller Seite aufgedrängte Abschlussprüfung der Sekundarstufe etablieren.

---

[58] Basic-Check und Multi-Check sind sich ähnlich. Im Falle von Basic-Check wird folgendes Ziel verfolgt: «Als normierte Prüfung für Schulabgänger bietet basic-check eine einheitliche Basis zur Vorabklärung von Fähigkeiten und Begabungen in Hinblick auf die zu wählende berufliche Grundbildung. Für Schüler wirkt der basic-check unterstützend für die Berufswahl. Für Lehrfirmen liefert er zusammen mit den Schulzeugnissen, dem Vorstellungsgespräch, der Analyse der Bewerbung und evtl. einer Schnupperlehre gute Anhaltspunkte für die Lehrlingsauswahl.» (vgl. Basic-Check 2004).

Begutachtet man die Qualität dieses Eignungstests anhand der zur Verfügung gestellten Beispielaufgaben (vgl. Basic-Check 2004) sowie der Interpretations-Anleitung (vgl. ebd.) genauer, so fallen einige gravierende Mängel auf:

- Es sind ausschliesslich Multiple-Choice-Aufgaben zu lösen (vgl. ebd.).
- Die Itemkonstruktion beschränkt sich meist auf die Stufen K1 (Kenntnisse) und K2 (Verstehen) der Bloomschen Lernzieltaxonomie. Gefordert ist das passive Erkennen isolierter, korrekter Antworten (vgl. Abbildung 2.9), weitergehende Fähigkeiten werden nicht erhoben (vgl. Basic-Check 2004).
- Wichtige Angaben zur Stichprobe wie die Geschlechterverteilung oder die regionale Verteilung fehlen: «Die Eichung erfolgte auf Grund einer grossen Zahl von Kandidatinnen und Kandidaten, welche diese Abklärung absolviert hatten.» (ebd.).
- Die Itemwahl wird nirgends begründet (vgl. ebd.), eine Differenzierung nach Kantonen und Lehrplänen fehlt.

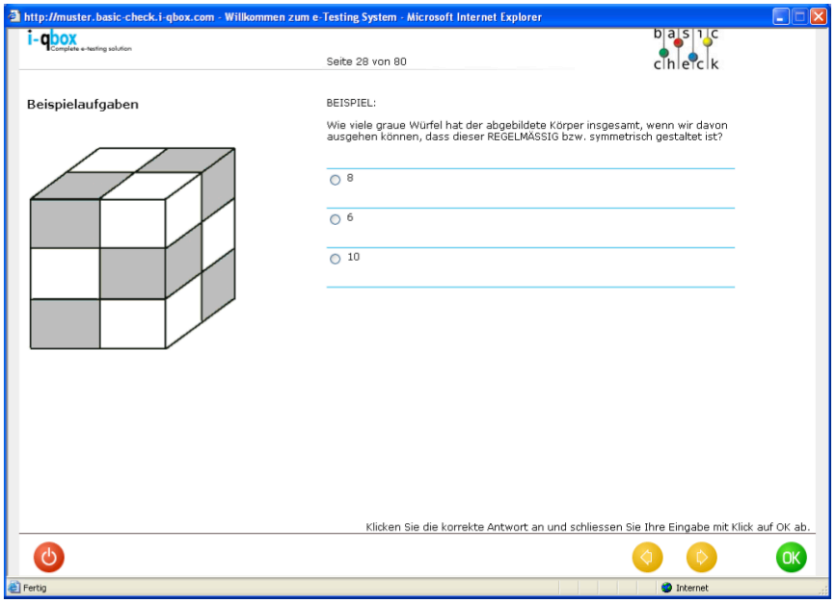

Abbildung 2.9: Beispielaufgabe «Denkaufgaben mit zwei- und dreidimensionalen Inhalten» (aus: Basic-Check 2004)

Hier von einem standardisierten und soliden Schulleistungstest zu sprechen, wäre verfehlt. Trotzdem: Lehrpersonen werden sich in Zukunft mit der Tendenz auseinandersetzen müssen, dass die Wirtschaft immer weniger auf Schulnoten und Zeugnisse vertraut und für immer mehr Firmen ein Bewerbungsdossier erst dann komplett ist, wenn die Resultate von Basic-Check oder Multi-Check beiliegen.

Die Schule ist also aufgefordert, für Firmen vergleichbarere Schülerbeurteilungen zu entwickeln. Ein in diese Richtung weisendes Projekt ist der seitens der Schweizerischen Konferenz der kantonalen Erziehungsdirektoren EDK vorgenommene kantonsübergreifende Harmonisierung HarmoS der obligatorischen Schule, oder einzelne kantonale Versuche zur Verbesserung der Beurteilungsqualität.[59] Dabei geht es aber nicht um die Frage, ob mit oder ohne Noten zu beurteilen sei. Viel wichtiger ist es, Möglichkeiten zu finden, damit Beurteilungen möglichst fair und verlässlich durchgeführt werden können und mit der Zeit auch Vergleiche über die eigene Schulklasse hinaus möglich sind (vgl. Rhyn 2002, p. 34). Es geht aber auch um die Frage, ob eine selektive Leistungsschule Absicht ist.

[59] Vgl. www.edk.ch/dyn/11659.php.

«Ist sie gewollt, verlangt sie möglichst eindeutige, zugleich faire Systeme der Beschreibung und Beurteilung von Leistungen, also weder des Charakters noch der Person, sondern im Blick auf das, was Lehrpersonen offenbar gut können, nämlich die Verteilung der individuellen Leistungsfähigkeit im Blick auf ihre Klasse» (Oelkers 2002, p. 15).

2.8 Grenzen der schulischen Leistungsdiagnostik

Erinnern wir uns an Frau Lutz, die sich überlegt, wie sie das Thema Berner Oberland in ihren Unterricht einbetten und am Ende prüfen will. Wir haben verschiedene Verfahren zur schulischen Leistungsdiagnostik vorgestellt. Diese Verfahren testen unterschiedliche Formen von Leistung (konvergente, divergente, sozial/emotionale). Zum einen erhält man mit dem Portfolio Informationen über die individuelle Entwicklung und Fortschritte der Lernenden, z.B. über das Verfassen von Texten, zum anderen wird mit standardisierten Schulleistungstests das Wissen der Schülerinnen und Schüler zu einem gewissen Zeitpunkt erhoben. Bei der Portfoliobeurteilung können auch neue Leistungsanforderungen wie selbstgesteuertes Lernen oder selbstverantwortliches Lernen eingelöst werden. Frau Lutz muss sich also überlegen, was sie von den Schülerinnen und Schülern am Ende der Themeneinheit erwartet. Möchte sie, dass sie alles Wichtige über das Berner Oberland wissen oder möchte sie, dass die Schülerinnen und Schüler lernen, wie man sich Informationen beschafft oder wie man ein Referat schreibt?

Die Lehrperson sollte sich auch Gedanken über die Gütekriterien (Objektivität, Validität und Reliabilität) machen, die im Hauptteil immer wieder angesprochen wurden. Im Abschnitt 2.3.1 zur Diagnosekompetenz wurde allerdings festgehalten, dass die traditionelle Leistungsmessung, z.B. mündliche Prüfungen oder Aufsatzprüfungen, den Kriterien der klassischen Testtheorie kaum oder nur in rudimentärer Form genügen. Da schon Objektivität nur selten gegeben ist, fehlen meistens auch Reliabilität und Validität. Hier muss man jedoch hinzufügen, dass es bei Lehrerinnen- und Lehrerurteilen *nicht immer* um die wissenschaftliche Genauigkeit geht, im Sinne von «je genauer, desto besser». Es spricht im Gegenteil einiges dafür, dass eine leicht positiv verzerrte Überschätzung des Leistungspotentials der Schülerinnen und Schüler pädagogisch günstig ist. Auch bei den weniger traditionellen Beurteilungsmethoden wie der Selbstbeurteilung oder der Portfoliobeurteilung sind Objektivität, Reliabilität und Validität nicht gegeben. Wenn man sich aber die Vorteile einer solchen Leistungsmessung vor Augen führt (z.B. Förderung der Mündigkeit des Lerners und der Fähigkeit zu eigenverantwortlichem Handeln bei der Selbstbeurteilung), so muss man sich überlegen, inwiefern man diese Fehlerquellen kontrollieren oder vielleicht sogar vernachlässigen kann. Im Gegensatz dazu sind bei standardisierten Schulleistungstests auf Grund ihrer fachgerechten Konstruktion und der Standardisierung der Durchführung, Auswertung und Interpretation sowohl Objektivität, Reliabilität als auch Validität gewährleistet. Sie erlauben auch Vergleiche über verschiedene Schulen oder verschiedene Regionen hinweg. Hier stellt sich aber vor allem die Frage, inwiefern ein standardisierter Test die spezifische Situation einer Klasse oder eines Schülers resp. einer Schülerin berücksichtigt, resp. inwieweit diese in der Interpretation der Ergebnisse berücksichtigt werden kann.

So scheint es also darauf hinauszulaufen, dass Beurteilungsformen, welche divergente und soziale Leistungen messen, nicht den Gütekriterien der Testtheorie gerecht werden, Beurteilungsformen, die rein auf das Erinnern von Informationen und auf den aktuellen Wissensstand abzielen, hingegen schon. Von der Schule wird aber beides verlangt. Die Schülerinnen und Schüler sollen Sachkompetenz erwerben, aber auch Sozial- und Selbstkompetenz. Also müssten am Ende einer Themeneinheit sowohl der aktuelle Wissensstand wie auch die individuellen Fortschritte gemessen werden. Da dies durch eine Beurteilungsform allein nicht erreicht werden kann, scheint es sinnvoll zu sein, mehrere Beurteilungsformen zu kombinieren, um den vielfältigen Anforderungen der pädagogischen Diagnostik gerecht werden.

3 Begabungsdiagnostik

3.1 Fallbeispiel

Yannis ist 10 Jahre alt, sehr feinfühlig und lebt mit seinen Eltern und einem um zwei Jahre älteren Bruder zusammen, der in Schule und Sport sehr erfolgreich ist. Yannis hingegen nervt seine Lehrerin, weil er ständig den Unterricht stört und seine Aggressionen nicht kontrollieren kann. In den Pausen ist er in Schlägereien verwickelt, was bei vielen Mitschülerinnen und Mitschülern Angst auslöst. Seine schulischen Leistungen bewegen sich im unteren Mittelmass; Yannis scheut Anstrengungen, sich zu verbessern. Weil er schliesslich auch zu Hause seinen Aggressionen freien Lauf lässt, nehmen die Eltern auf Anraten der Lehrerin schulpsychologische Beratung in Anspruch. Eine Begabungsabklärung deckt auf, dass Yannis in der Schule weit unter seinem Potenzial arbeitet und eigentlich ebenso gute Leistungen wie sein Bruder erbringen sollte. Im mathematischen Bereich ist er Gleichaltrigen um etwa drei Jahre voraus. Nach Absprache aller Beteiligten kann Yannis in eine andere Schule wechseln, wo der Lehrer eine Mehrjahrgangsklasse unterrichtet und stark individualisierend arbeitet. Yannis konnte von der Schulpsychologin nur durch aufwändige Gespräche vom Sinn dieser Versetzung überzeugt werden. Zudem kann er nun einen Nachmittag pro Woche der Klasse fernbleiben und Förderstunden in Mathematik besuchen. Dies tut er anfänglich nur widerwillig, nach den ersten Erfahrungen aber wächst sein Interesse deutlich. In der neuen Klasse arbeitet er bald mit den fortgeschrittensten Schülerinnen und Schülern zusammen und findet dort einen Freund. Nach einem Jahr tritt er, ein Jahr jünger als die anderen, ins Gymnasium über. Noch ist Yannis nicht der brave Junge, aber sein destruktives Verhalten hat sich sichtbar gemildert (vgl. Huser 1999, p. 24f.).

Wie das Beispiel von Yannis zeigt, sind diagnostische Entscheidungen im Alltag von Eltern und Lehrpersonen stets präsent und ziehen oft Konsequenzen nach sich, die das Leben prägen (vgl. Ingenkamp 1997, p. 155). Wenn es um den schulischen Erfolg einer Schülerin oder eines Schülers geht, beschäftigen sich Eltern und Lehrpersonen u.a. mit Fragen wie:

- Hat das Kind ausserordentliche Neigungen und soll es deshalb speziell gefördert werden?
- Ist das Kind in der regulären Schule überfordert und braucht es daher eine Stütze im sonderpädagogischen Bereich?

Damit die aufgeworfenen Fragen beantwortet werden können und eine entsprechende Entscheidung getroffen werden kann, sollte die Persönlichkeit des Kindes sowie seine Lebensgeschichte mit in Betracht gezogen werden. Dabei geht es vor allem um das Erkennen von «inneren» Merkmalen wie z.B. Intelligenz, Begabungen, Interessen und Kenntnissen und in diesem Kapitel besonders um die Diagnostik von Hochbegabung.[60]

Zunächst werden die Problematik des Schulerfolgs bei Hochbegabten und das Thema der Intelligenz als Kernbereich der Hochbegabung aufgegriffen (Abschnitt 3.1.1 und Abschnitt 3.1.2). Im Anschluss wird die Diagnostik der Hochbegabung behandelt: Es werden sowohl auf die standardisierten als auch auf die nicht-standardisierten Verfahren eingegangen, wobei sich der Fokus auf die Intelligenzdiagnostik richtet (Abschnitt 3.2). Im Abschnitt 3.3 wird einer der meist verwendeten

[60] Da die «Hochbegabung» ein nicht direkt beobachtbares Merkmal ist, erstaunt es wenig, dass in der Hochbegabungsdiskussion bereits weit über hundert Definitionen vorherrschen, bei der fast jede irgendwie plausibel erscheint. Bezeichnungen wie «Höchstbegabung», «höhere Begabung», «Hochbefähigte», «Hochintelligente», «Spitzenbegabte», «besonders Befähigte» machen nicht klar, wie die Begriffsnuancen voneinander zu unterscheiden sind. Mehrheitlich werden diese Begriffe heute bedeutungsgleich mit «Hochbegabung» verwendet (vgl. Stamm 1992, p. 55ff.).

Intelligenztests, der HAWIK-III, vorgestellt und die Anwendungsmöglichkeiten von Intelligenztests bei Lernstörungen diskutiert (Abschnitt 3.4). Das Kapitel wird mit der Kritik am Einsatz von Intelligenztests (Abschnitt 3.5) und einem Fazit (Abschnitt 3.6) abgeschlossen.

3.1.1 Hochbegabung und Schulerfolg

Wie das eingangs erläuterte Fallbeispiel verdeutlicht, können gerade hochbegabte Schülerinnen und Schüler im schulischen Bereich Schwierigkeiten haben. Wie andere Lernende auch brauchen Hochbegabte für eine optimale Entwicklung ihrer Fähigkeiten Anregung und Ermutigung durch die Umwelt. Hochbegabung ist kein statisches Gebilde, sondern entfaltet sich oder verkümmert unter den entsprechenden Einflüssen der Umwelt. Es ist entscheidend, dass eine besondere Begabung erkannt und gefördert wird, die Schülerin oder der Schüler auf Verständnis stösst und in ihrer oder seiner Gesamtpersönlichkeit adäquat betreut werden kann.

Hochbegabte Lernende zeichnen sich meist durch sehr früh entwickelte, weit überdurchschnittliche Fähigkeiten und Interessen aus, durch die sie Gleichaltrigen oft beträchtlich voraus sind. Dies kann den logisch-mathematischen, den sprachlichen, den musikalischen, den bildnerisch-künstlerischen, den sportlichen oder den sozialen Bereich – manchmal auch mehrere dieser Bereiche gleichzeitig – betreffen. Zu beachten ist, dass nicht zwangsläufig alle Fähigkeiten weit überdurchschnittlich ausgeprägt sein müssen (vgl. Stamm 1992, p. 312f.). Hochbegabung ist also ein *Interaktionsprodukt zwischen individueller Anlage und sozialer Umgebung*.

Bleiben entsprechende Massnahmen aus, kann es zum Phänomen des *Underachievers* kommen. Underachiever bzw. Minderleisterinnen und -leister sind Schülerinnen und Schüler, die unabhängig ihres IQs[61] mehr leisten könnten, als sie tatsächlich an Leistung erbringen. Es besteht somit eine Diskrepanz zwischen dem Potenzial der Lernenden und dem, was sie zeigen. Überträgt man diese Thematik auf den hochbegabten Yannis, muss festgestellt werden, dass er in der Schule nicht in konstruktiver Weise auf seine Hochbegabung aufmerksam machte. Underachievement kann einerseits durch schulische Unterforderung, durch Anpassung an die durchschnittlichen Erwartungen, andererseits aber auch durch Überforderung, übertriebene schulische oder elterliche Erwartungen bedingt sein. Motivationale Aspekte scheinen dabei eine überragende Bedeutung zu spielen.[62] So sollten nebst einer Anleitung zu Arbeitstechniken und -rhythmen attraktive schulische oder ausserschulische Angebote gemacht werden, bei welchen die Lernenden ihre Leistungsfähigkeiten bewusst wahrnehmen lernen und erkennen, dass Anstrengung zum Erfolg führen kann.

In der pädagogischen Diagnostik nimmt die *Intelligenzmessung* trotz der vielen Gesichter der Hochbegabung eine zentrale Rolle ein.

---

[61] Stern entwickelte 1912 den Intelligenzquotienten IQ, bei dem das Intelligenzalter durch das Lebensalter geteilt wird. Dadurch hatte er eine Masszahl für Intelligenz gefunden, anhand derer Kinder unterschiedlichen Alters bezüglich ihrer intellektuellen Fähigkeiten miteinander verglichen werden konnten. Spätere Forscher multiplizierten das Ergebnis – eine Bruchzahl – dann mit dem Faktor 100, um Kommazahlen zu vermeiden. Ein Intelligenzquotient von 100 bedeutet, dass eine Übereinstimmung von Lebensalter und Intelligenzalter vorliegt und kennzeichnet damit den durchschnittlichen IQ. Für Erwachsene jedoch lässt sich nach dieser Formel die Ausprägung der Intelligenz nicht mehr sinnvoll berechnen, da das Lebensalter immer weiter steigt, während das Intelligenzalter ab einem gewissen Zeitpunkt konstant bleibt. Wechsler hat dieses Problem beseitigt, indem er 1955 den sogenannten Abweichungs-IQ entwickelte. Dabei handelt es sich um eine Formel, in die das individuelle Testergebnis einer Person sowie Kennzahlen ihrer Altersgruppe eingesetzt werden. Dieser Wert stellt dar, was man heute (begrifflich nicht mehr korrekt) als IQ bezeichnet. Auf diese Weise lassen sich die Leistungen von Kindern und Erwachsenen unterschiedlichen Alters vergleichen (vgl. Holling & Kanning 1999, p. 25).

[62] Vgl. auch Huser (1999, p. 28); Hanses & Rost (1998, p. 55); Stamm (1992, p. 318).

3.1.2 Intelligenz als Kernbereich der Hochbegabung

Wenn von *Intelligenztests* bzw. *Intelligenzdiagnostik* die Rede ist, besteht stets eine untrennbare Verbindung mit der Definition von Intelligenz. Die menschliche Intelligenz ist eines der im 20. Jahrhundert am intensivsten erforschten Konstrukte (vgl. Abschnitt 1.1.1). Trotz unzähliger Untersuchungen und Entwicklungen von Intelligenztests konnten sich die Forschende bisher nicht einigen, was Intelligenz sei.

Der Wert aller hervorgebrachten Definitionen ist dadurch begrenzt, dass sie unterschiedlich interpretiert werden können. Spearman sagte dazu treffend: «Intelligenz ist ein Wort mit so vielen Bedeutungen, dass es zuletzt überhaupt keine mehr hat» (Spearman 1927, o. S., zit. n. Ingenkamp 1997, p. 172). Daher bevorzugt man immer häufiger operationale Definitionen, indem man durch Testaufgaben definiert, was in diesem Zusammenhang als Intelligenz verstanden wird. Das Konstrukt Intelligenz wird damit ganz an bestimmte Tests gebunden. In diesem Sinne hat Boring schon 1923 teils ironisierend, teils ernüchtert formuliert: «Intelligenz ist das, was der Intelligenztest misst» (Boring 1923, o.S., zit. n. ebd., p. 172f.). Eine solche Auffassung verspricht immerhin nichts Falsches (vgl. Ingenkamp 1997, p. 172f.).

In den zahlreichen Definitionen der Intelligenz werden in jeweils unterschiedlicher Breite und Gewichtung eine Vielzahl von geistigen, musisch-künstlerischen, psycho-motorischen, sozialen und emotionalen Fähigkeiten berücksichtigt. Da die *intellektuellen Fähigkeiten* meist im Mittelpunkt der Betrachtung stehen, gehören Intelligenztests zu den mit Abstand am häufigsten eingesetzten Verfahren (vgl. Holling & Kanning 1999, p. 21). Bei der Intelligenzmessung hat sich für die Definition von Hochbegabung die Grenze von IQ = 130 durchgesetzt. Personen, die oberhalb dieses Wertes liegen, werden als hochbegabt bezeichnet. Aufgrund dieses Grenzwerts, der lediglich eine Konvention, also eine Abmachung darstellt, werden 2,1% der Bevölkerung als hochbegabt klassifiziert. Stellt man die relative Häufigkeit der Intelligenzquotienten in einer Bevölkerung grafisch dar, erhält man eine glockenförmige Kurve, die sogenannte *Normalverteilungskurve*.

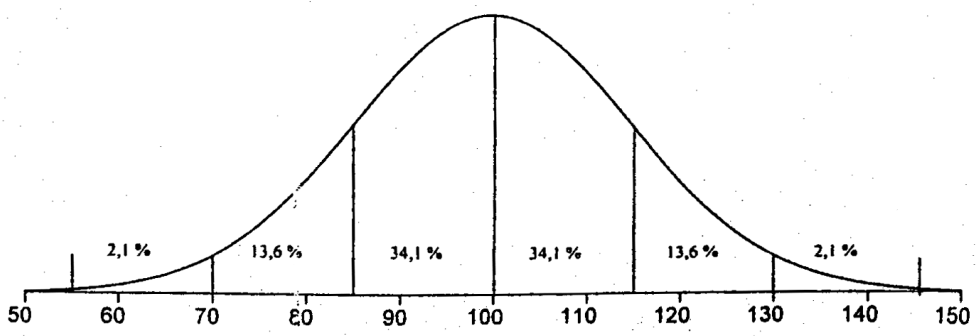

Abbildung 3.1: Normalverteilung der Intelligenzquotienten (aus: Holling & Kanning 1999, p. 23)

Die abgeteilten Flächen unter der Kurve zwischen der horizontalen und den vertikalen Achsen verdeutlichen die prozentuale Häufigkeit, mit der die jeweiligen Wertebereiche auftreten. So ist, wie oben bereits erwähnt, zu erkennen, dass nur jeweils rund zwei Prozent der Bevölkerung einen sehr niedrigen IQ (unter 70) oder einen sehr hohen IQ (über 130) haben und – im zweiten Falle – als hochbegabt bezeichnet werden. Im Kapitel 1.5 wurde bereits auf die Normalverteilung der IQ-Werte eingegangen.

3.2 Diagnostik der Hochbegabung

Die Diagnostik ist stark mit der Definition und den entsprechenden *Modellen der Hochbegabung* verbunden. Wenn in einem Modell beispielsweise als einziges Unterscheidungskriterium zwischen Hochbegabten und Normalbegabten ein bestimmter Intelligenzwert festgelegt ist, so wird sich die Identifikation von Hochbegabten darauf beschränken, jene Schülerinnen und Schüler herauszufiltern, die diese Norm erfüllen. Modelle hingegen, die einen multifaktoriellen Ansatz verfolgen, erfordern eine breiter angelegte Diagnostik. In diesem Fall werden über die Intelligenz hinaus auch *kreative Fähigkeiten* und *motivationale Komponenten* erhoben. Diese Merkmale sind jedoch schwieriger zu erfassen als Intelligenz. Da auch in komplexeren Modellen zur intellektuellen Hochbegabung wie etwa dem Triadischen Interdependenzmodell oder dem Münchner Begabungsmodell die kognitiven Fähigkeiten den Dreh- und Angelpunkt darstellen, werden diese bei der Diagnostik von intellektueller Hochbegabung häufig am stärksten fokussiert. Damit stehen Intelligenz- und Leistungstests nach wie vor im Vordergrund (vgl. Holling & Kanning 1999, p. 21).

Generell unterscheidet man zwischen den *standardisierten* (formellen) und den *nicht-standardisierten* (informellen) *Verfahren* in der Diagnostik der Hochbegabung.

Im folgenden Kapitel werden die informellen Verfahren anhand des Lehrerinnen- und Lehrerurteils kurz betrachtet. Im Anschluss wird etwas ausführlicher auf die standardisierten Verfahren eingegangen. Der grösste Vorteil dieser liegt darin, dass sie vergleichbare Daten für alle Lernenden liefern können, was nicht-standardisierte Verfahren nicht zu leisten imstande sind.

3.2.1 Nicht-standardisierte Verfahren

Neben Elternurteil und Peerbeurteilung gehören auch *Lehrerinnen- und Lehrerurteile* zum Bereich der nicht-standardisierten Verfahren in der Diagnostik von Hochbegabung. Wie zutreffend die Potenzialeinschätzung im Leistungsbereich einer Schülerin oder eines Schülers sein kann, wenn sich diese ausschliesslich auf das Lehrerinnen- und Lehrerurteil bezieht, zeigt das Beispiel von Yannis. Nicht nur er kommt mit seiner – noch unbemerkten – Hochbegabung nicht klar, auch die Lehrerin schätzt das Potenzial des Jungen nicht richtig ein. Underachiever Yannis kann von Glück sprechen, dass seine Lehrperson den schulpsychologischen Dienst in Anspruch genommen hat und somit weitere Schritte in die Wege geleitet werden konnten.

Über den Wert der Datenquelle «Lehrerinnen- und Lehrerurteil» existieren zahlreiche Untersuchungen, welche zum Teil widersprüchliche Ergebnisse liefern. Aufgrund der vielfältigen Erfahrung und der breiten Vergleichsbasis wird von der Lehrperson erwartet, dass sie relativ gut die allgemeine Leistungsfähigkeit des Schülers oder der Schülerin beurteilen und über spezifische Fähigkeiten und Schwächen ihrer Klasse(n) Auskunft geben kann. Eine Studie zeigte jedoch, dass Nominationen durch Lehrer und Lehrerinnen nicht sehr aussagekräftig sind: viele Lernende benannten sie als hochbegabt, welche sich im Nachhinein nicht als solche erwiesen und übersahen andere, welche tatsächlich hochbegabt waren. Des Weiteren urteilen sie oft nach *Stereotypen*, da sie – was Hochbegabung betrifft – über unzutreffende Vorstellungen verfügen. Mädchen wie auch ausländische Schülerinnen und Schüler oder solche aus armen und bildungsfernen Familien werden oft unterschätzt. Nicht zuletzt orientieren sich Lehrpersonen stark an den Schulnoten, die eher ein Abbild momentaner Leistungen sind, als dass sie Potenzial aufzuzeigen vermögen. Fleissigen Schülerinnen oder Schülern, die zudem fähig und bereit sind, sich den Erwartungen der Lehrerin oder des Lehrers anzupassen, wird oft eine intellektuelle Begabung zugeschrieben. Bringt eine Schülerin oder ein Schüler diese Eigenschaften nicht mit, wird dies mit mangelnder Begabung verwechselt (vgl. Ey-Ehlers 2001, p. 87f.).

Wild (1991, p. 150f.) hingegen spricht von einer relativ hohen durchschnittlichen Übereinstimmung von Lehrerinnen- und Lehrerurteilen und Testwerten aus Intelligenztests. Dass unter diesen Perso-

nenurteilen dennoch eine breite Streuung vorliegt, führt er auf die unterschiedliche diagnostische Kompetenz der Lehrpersonen zurück. Dass sich diese über die Zeit hinweg nicht verbessern lasse, findet er problematisch. Lehrerinnen und Lehrer, die während zwei bis drei Schuljahren ihre Klasse(n) kennen lernen, beurteilen diese nicht adäquater als Lehrpersonen, die mit ihren Schülerinnen und Schülern erst seit einem halben Jahr zu tun haben.

Nicht-standardisierte Verfahren wie das Lehrerinnen- und Lehrerurteil weisen zum Teil also erhebliche Mängel in der Diagnostik von Hochbegabung auf. Im Falle von Yannis konnte erst ein Intelligenztest helfen, diese zu identifizieren. Daher wird nach einigen allgemeinen Bemerkungen zu Testmethoden im Weiteren nur auf standardisierte Verfahren eingegangen.

3.2.2 Standardisierte Verfahren

Wenn wir umgangssprachlich von «Test» sprechen, kann damit *jede Art von Probe oder Prüfverfahren* gemeint sein. Eine für die pädagogische Praxis relevantere Definition schlägt Ingenkamp (1997) vor, in die er auch die Gütekriterien von Tests mit einbezieht:

> «Tests sind Verfahren der Pädagogischen Diagnostik, mit deren Hilfe eine Verhaltensstichprobe, die Voraussetzungen für oder Ergebnisse von Lernprozessen repräsentieren soll, möglichst vergleichbar, objektiv, zuverlässig und gültig gemessen und durch Lehrer oder Erzieher ausgewertet, interpretiert und für ihr pädagogisches Handeln nutzbar gemacht werden kann» (ebd., p. 80).

Zu beachten ist, dass Tests nur ein Verfahren darstellen, welches uns helfen kann, den Beurteilungsprozess zu professionalisieren.

3.2.2.1 Kreativitätstest

Kreativitätstests sind im Allgemeinen umstritten und werden eher selten eingesetzt, nicht zuletzt auch wegen ihrer geringen Bewährung und der schwierigen Auswertung (vgl. Rost 1989, p. 14). Sie beschränken sich auf die Erfassung des divergenten Denkens als einem Aspekt der Kreativität. Allgemein werden vier verschiedene Herangehensweisen unterschieden, wobei die Biographische Methode[63] im deutschen Sprachraum nicht existiert (vgl. Lukesch 1998, p. 285f.).

3.2.2.2 Selbstbeurteilungstests

Hierbei wird entweder versucht, ein spezielles Interessens- und Tätigkeitsprofil oder die Selbstsicht in Bezug auf kreativitätsspezifische Merkmale zu erfassen. Für letztere Vorgehensweise existiert eine deutsche Version des «Group Inventory for Finding Creative Talent» (GIFT) durch Urban (1982). Mit diesem Instrument können Kinder als kreativ identifiziert werden, die ev. von Lehrpersonen als wenig begabt und störend charakterisiert werden.

3.2.2.3 Fremdbeurteilungstests

Die Urteile können sich dabei entweder auf Produkte oder auf Personen beziehen. Als Beurteilende können Eltern, Lehrpersonen oder Mitschülerinnen und -schüler beigezogen werden. Für den deutschen Sprachraum liegt hierzu ein Verfahren für die Einschätzungen von Lehrpersonen sowie von Mitschülerinnen und -schüler von Seiffge-Krenke (1974) vor.

---

[63] Im Sinne eines biographischen Vorgehens kann bereits erbrachte kreative Produktion erfasst werden (u.a. Anzahl bisheriger Publikationen, Erfindungen, Patente, Preise bei Wettbewerben) (vgl. Lukesch 1998, p. 285).

3.2.2.4 Psychometrische Verfahren

Diese Methoden wurden letztlich durch Guilford (1950) initiiert. Typisch für diese Verfahren ist, dass ohne Zeitdruck eine Vielzahl von Lösungen, die unvorhersehbar waren, angeregt werden sollen.

3.2.2.5 Einzelintelligenztest

In fast allen Modellen wird die *kognitive Leistungsfähigkeit* als zentraler Bestandteil von Hochbegabung genannt, so dass die Verwendung von Intelligenztests zu den wichtigsten Verfahren zählt. Abgesehen von wenigen Ausnahmen konzipieren fast alle der im deutschen Sprachraum verwendeten Intelligenztests Intelligenz als Ensemble von Fähigkeiten, wobei deren *Gesamtergebnis sich aus der Summe der Werte in Teilbereichen einer Aufgabensammlung zusammensetzt*. Die Teilbereiche einer solchen Aufgabensammlung berücksichtigen ihrerseits verschiedene Fähigkeiten oder *Intelligenzfaktoren*, die einem entsprechenden Intelligenzmodell zugrunde liegen (vgl. Stamm 1992, p. 91). Die Intelligenztests sind ein zentraler Bestandteil der Intelligenzdiagnostik. Lukesch zählt zahlreiche Einzelintelligenztests[64] auf (1998, p. 249ff.):

3.2.2.6 Eindimensionale Intelligenztests

Ein *eindimensionaler Intelligenztest* beruht auf Messungen in einem spezifischen Teilleistungsbereich. Zu solchen Tests gehört beispielsweise der sprachliche Leistungstest (SASKA), der nur aus einem Verbalteil besteht.

3.2.2.7 Mehrdimensionale Intelligenztests

Ein *mehrdimensionaler Intelligenztest* besteht aus Messverfahren in mehreren spezifischen Teilleistungsbereichen. Zu solchen Tests gehört beispielsweise der «Hamburger-Wechsler- Intelligenztest für Kinder» (HAWIK-III), der sich aus einer Kombination von einem *Verbal- sowie einem Handlungsteil* zusammensetzt. Da dieser Test das im deutschen Sprachraum am meisten verwendete Verfahren ist, wird er im folgenden Kapitel ausführlicher vorgestellt.

3.3 Intelligenztest: HAWIK-III

Die Wechsler-Tests gehören zu den prominentesten und empirisch am intensivsten untersuchten Intelligenzverfahren und haben mit ihrer langen Tradition einen international beachteten Standard in der Intelligenzdiagnostik festgelegt. Der Hamburger-Wechsler-Intelligenztest für Kinder (HAWIK-III) stellt eine Art «Goldstandard» in der Intelligenzdiagnostik im deutschsprachigen Raum dar. Er wird sowohl im Bereich der Schulpsychologie als auch der Kinder- und Jugendpsychiatrie verwendet. Der HAWIK-III wird vor allem eingesetzt bei Schulversagen, Feststellung des allgemeinen Leistungsniveaus, Legasthenie, Dyskalkulie sowie allen Formen von Teilleistungsschwächen und Über- oder Unterforderung. Beim HAWIK-III handelt es sich um die 1999 erschienene 3. Auflage[65] als Übersetzung und Adaptation des bereits 1991 in den USA veröffentlichten WISC III (Wechsler Intelligence Scale for Children) von Wechsler. Der HAWIK-III wurde gegenüber dem HAWIK-R (1983) um zwei Untertests erweitert.

---

[64] Der Einzelintelligenztest – im Vergleich zu einem in einer Gruppe geführten Intelligenztest – ist das Kernverfahren der Intelligenzmessung.

[65] Diese 3. Fassung wurde von Uwe Tewes (Medizinische Hochschule Hannover), Peter Rossmann (Universität Graz) und Urs Schallberger (Universität Zürich) erstmals gemeinsam für Deutschland, Österreich und die deutschsprachige Schweiz normiert.

Der HAWIK-III ist ein mehrdimensionaler Test, der auf dem *Zweifaktorenmodell* von Spearman basiert. Als Intelligenztest der Einzelfalldiagnostik untersucht er den Stand der kognitiven Entwicklung von Kindern und Jugendlichen im Alter von 6 bis 16 Jahren.[66] Der etwa 60 bis 90 Minuten dauernde Test ist ein *Niveautest*, bei dem es herauszufinden gilt, wie viele Aufgaben die Lernenden in vorgegebener Zeit lösen können. Alle Aufgaben sind nach aufsteigender Schwierigkeit angeordnet. Die allgemeine Intelligenz wird im Sinne des Intelligenzkonzepts von Wechsler in den beiden spezifischen Untertests *Verbal- und Handlungsteil* erfasst, was die Darstellung eines *Leistungsprofils* ermöglicht. Das Ergebnis des Verbalteils bezeichnet man häufig als *sprachliche oder verbale Intelligenz*, das Ergebnis des Handlungsteils als *praktische oder handlungsbezogene Intelligenz*. Diese begriffliche Unterscheidung ist zwar nicht ganz eindeutig und korrekt, kann aber als pragmatische Unterscheidung vernünftig erscheinen. Auf diese Weise kann geprüft werden, ob die Gesamtleistung in erster Linie auf gute sprachliche Begabungen oder eher auf gute praktische Begabungen zurückzuführen ist bzw. ob beide Aspekte etwa gleich stark ausgeprägt sind. Der Gesamttest gliedert sich in die genannten zwei Teile mit je 5 Untertests und dem Zusatzteil mit 3 Untertests:

Verbalteil (VT)	*Handlungsteil (HT)*	*Zusatzteil*
Allgemeines Wissen (AW)	Bilderergänzen (BE)	Zahlennachsprechen (ZN) für den VT
Gemeinsamkeiten finden (GF)	Zahlensymboltest (ZS)	Labyrinth-Test (LT) und Symbolsuche (SS) für den HT
Rechnerisches Denken (RD)	Bilderordnen (BO)	
Wortschatz-Test (WT)	Mosaik-Test (MT)	
Allgemeines Verständnis (AV)	Figurenlegen (FL)	

Diese 13 Untertests werden auf die vier Teilleistungsbereiche Sprachliches Verständnis, Unablenkbarkeit, Wahrnehmungsorganisation und Arbeitsgeschwindigkeit aufgeteilt.

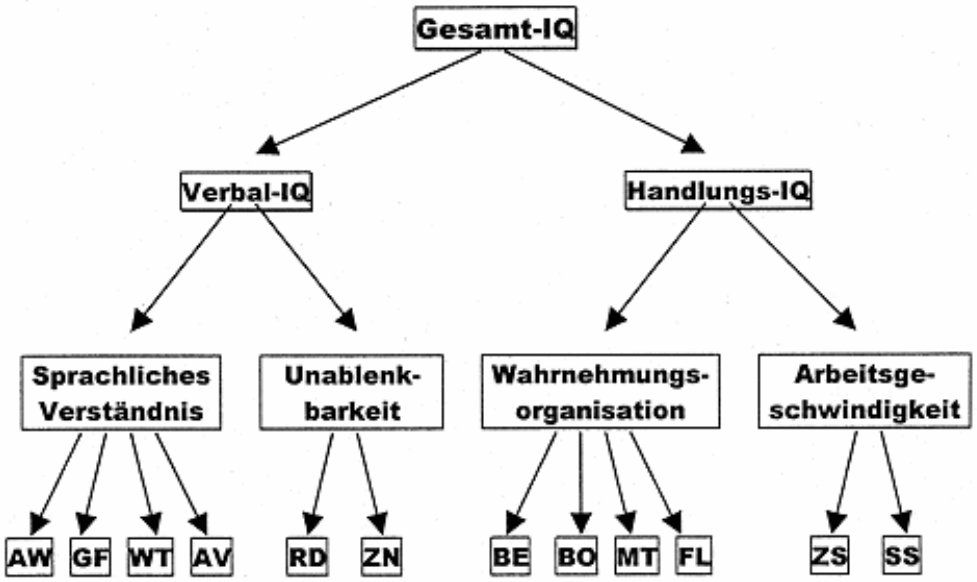

Abbildung 3.2: Übersichtsmodell des HAWIK-III Aufbaus (aus: Testzentrale der Schweizer Psychologen: www.testzentrale.ch)

[66] Der Test ist für Jugendliche bis zum Alter von 16 Jahren und 11 Monaten gedacht, da die Normen jeweils separat für Altersstufen im Abstand von 3 Monaten (und geschlechtspezifisch) vorliegen. Für Kinder unter 6 Jahren gibt es den HAWIVA- Test und für Jugendliche ab 17 Jahren gibt es den HAWIE-Test.

Jeder Untertest im Verbal- und im Handlungsteil beinhaltet folgende Unterscheidungskriterien:

- Gesamtzahl der Items;
- Maximale Punktzahl pro Items;
- Einstiegsstufen[67];
- Umkehrregel[68];
- Abbruchkriterium[69].

Im folgenden Kapitel werden der Verbal- und der Handlungsteil mit den dazugehörigen Untertests näher erläutert.

3.3.1 Verbalteil

Der Verbalteil prüft die sprachliche Intelligenz auf Grund des *sprachlichen Verständnisses*. Dazu gehören sowohl die sprachlichen Kenntnisse und das verbale Ausdrucksvermögen durch Beantwortung von mündlich gestellten Fragen zur Prüfung des Allgemeinwissens, der Wortbedeutungen und des schlussfolgernden Denkens, als auch die Fähigkeit, die eigenen Gedanken in Worte zu fassen. Der Verbalteil besteht aus folgenden Untertests:

Allgemeines Wissen (AW): Es werden 33 mündliche Fragen gestellt und die Lernenden sollen die dazu passende Antwort geben.[70] Damit wird der Wissensumfang der Schülerin oder des Schülers ermittelt (vgl. Tabelle 3.1).

---

[67] Die Einstiegstufe bedeutet diejenige Aufgabe, mit der eine Aufgabenserie eines Untertests für ein Kind in einem bestimmten Alter beginnt.

[68] Die Umkehrregel wird nach Art der Testaufgabe festgelegt. Beispielsweise lautet die Umkehrregel beim Untertest Rechnerisches Denken folgendermassen: Bei den Altersgruppen 7-16 werden die Aufgaben unterhalb der Startaufgabe in rückwärtiger Folge vorgegeben, wenn eine der beiden zuerst vorzugebenden Aufgaben nicht oder falsch gelöst wird (vgl. Tewes, Rossmann & Schallberger 2000a, p. 68).

[69] Das Abbruchkriterium bedeutet die Anzahl der Fehler in einer Aufgabenfolge, nach der die Durchführung des Untertests abgebrochen wird.

[70] Wenn drei bis fünf aufeinander folgende Aufgaben nicht gelöst bzw. falsch gelöst wurden, wird der Untertest abgebrochen.

Tabelle 3.1: Beispiele aus dem Untertest Allgemeines Wissen (aus: Tewes, Rossmann & Schallberger 2000a, p. 127f.)

Testfrage	**gültige Antwort**
Wie viele Beine hat ein Hund?	Vier
Was musst Du machen, damit das Wasser kocht?	Erhitzen... den Wasserkessel füllen und den Herd anstellen... es auf den Herd stellen... Feuer darunter anmachen... es in die Mikrowelle stellen... die Herdplatte anstellen (oder irgendeine andere Antwort, aus der hervorgeht, dass das Wasser erhitzt werden muss).
Welcher Tag kommt nach dem Donnerstag?	Freitag
Wer war Christoph Kolumbus? (**Schweiz:** Wer ist Christoph Kolumbus gewesen?)	Ein Forscher... Einer von denen, die Amerika entdeckt haben... Ein Kapitän, der Amerika entdeckte... Er wurde von König Ferdinand und Königin Isabella von Spanien beauftragt, einen neuen Seeweg für den Handel mit Gewürzen zu entdecken...
In welcher Himmelsrichtung geht die Sonne unter?	Im Westen (Falls das Kind in die entsprechende Himmelsrichtung zeigt, fragt man: „Ja, aber welche Himmelsrichtung ist das?“).
Wie heisst die Hauptstadt von Griechenland?	Athen
Warum donnert es beim Gewitter?	Weil durch den Blitz die Luft in Bewegung versetzt wird... Wegen der plötzlichen Luftbewegung... Die elektrischen Entladungen des Blitzes verursachen den Donner...
Was ist ein Barometer?	Es zeigt den Luftdruck an... Es misst den atmosphärischen Druck... Es sagt Veränderungen des Wetters voraus... Es liefert einen Wetterbericht... Es zeigt an, ob es Regen gibt... Ein Druckanzeiger (Druckmesser).
Wie gross ist der Erdumfang?	Jede Antwort zwischen 40 000 und 42 000 km

Gemeinsamkeiten finden (GF): Bei 25 Wortpaaren soll das Kind jeweils angeben, was das Gemeinsame dieser Begriffe ist. Auf diese Weise prüft der Test die Abstraktionsfähigkeit und die Kategorienbildung von zwei mündlich genannten Begriffen. Beispiele: «Säge – Kneifzange», «Baum – Gras».

Wortschatz –Test (WT): Damit wird der kindliche Wortschatz geprüft, indem die Lernenden 44 Begriffe erklären sollen. Beispiele: «Brot», «immun».

Allgemeines Verständnis (AV): Mit 20 Fragen werden die Fähigkeit zur ethischen Wertung, die Fähigkeit, Alltagssituationen zu erfassen und zu beurteilen und lebenspraktische Sprachverständnisse zu überprüfen, ermittelt. Beispiel: «Warum sehen Gegenstände und Personen in der Ferne kleiner aus?».

Rechnerisches Denken (RD): Anhand von 29 Aufgaben wird die Rechenfähigkeit der Lernenden ermittelt. Es wird die Fähigkeit geprüft, einfache Zahlenoperationen unter Verwendung der Grund-

rechenarten im Kopf durchführen zu können sowie die geistige Wendigkeit und schlussfolgerndes Denken bei der Umsetzung alltäglicher Aufgaben in mathematischen Lösungsschritten.

Beispiele:

- Fritz hat 4 Bonbons, und seine Mutter gibt ihm 2 Bonbons dazu. Wie viele Bonbons hat er jetzt im Ganzen?
- Du hast 10 Stücke Schokolade und isst 3 davon. Wie viele Stücke Schokolade hast du dann noch?
- Ein Mann hat 12 Zeitungen und verkauft 5 davon. Wie viele Zeitungen hat er noch?
- Philipp ist 36 km gewandert. Pro Stunde ist er 4 km gegangen. Wie viele Stunden hat er für die ganze Wanderung gebraucht?
- Marion wandert doppelt so weit wie Carola. Marion wandert 17 km. Wie viele km wandert Carola?
- Ein Autofahrer macht nach 28 km eine Pause. Das sind zwei Drittel seiner Gesamtstrecke. Wie viele km beträgt die Gesamtstrecke?
- In einer Schulklasse sind 38 Kinder. Es sind 12 Schüler mehr als Schülerinnen. Wie viele Schülerinnen sind in der Klasse?

Zahlennachsprechen (ZN): Dieser fakultative Test prüft das Kurzzeitgedächtnis. Die Schülerin oder der Schüler soll sieben (maximal neunstellige) Zahlenreihen vorwärts und sieben (maximal achtstellige) Zahlenreihen rückwärts wiedergeben können.

3.3.2 Handlungsteil

Der Handlungsteil prüft die praktische Intelligenz auf Grund des *Wahrnehmungsverständnisses*. Dazu gehören die nonverbalen Denkvorgänge, die Integration visueller Reize, das nonverbale Schlussfolgern und die Fähigkeiten des räumlichen Denkens. Zugleich werden auch die Denkgeschwindigkeit und die motorische Geschwindigkeit bei der Bearbeitung verschiedener nonverbaler Aufgaben mit geringen Schwierigkeitskomponenten getestet.

Der Handlungsteil besteht aus folgenden Untertests:

Bilderergänzen (BE): Den Lernenden werden nacheinander 33 Bilder vorgelegt, auf denen ein wichtiger Teil fehlt, der erkannt und benannt werden soll. Dadurch werden die visuelle Auffassungsgabe, die Differenzierung zwischen wesentlichen und unwesentlichen Elementen, die Realitätskontrolle und die geistige Beweglichkeit geprüft (vgl. Abbildung 3.3).

Bilderordnen (BO): Die Kinder sollen die Bildkarten mehrerer Bildserien in die richtige Reihenfolge bringen, so dass sich eine sinnvolle Bildergeschichte ergibt. Dadurch wird einerseits die Fähigkeit, Bilder in eine zeitlich-logische Reihenfolge zu legen und anderseits die Fähigkeit, dabei komplexe Alltagssituationen in ihren Abläufen zu erfassen, geprüft (vgl. Abbildung 3.4).

Mosaik-Test (MT): Der Schülerin, dem Schüler werden 17 Vorlagen präsentiert, die sie, er mit vier bis neun identisch gemusterten Würfeln nach Abbildung genau legen soll, wobei die Abweichung der Lage des gesamten Musters vom Original 30° nicht überschreiten darf. Mit diesem Untertest werden Kombinationsfähigkeit, räumliches Vorstellungsvermögen und visuell-motorische Koordination geprüft.

Abbildung 3.3: Beispiel aus dem Untertest Bilderergänzen (aus: Tewes, Rossmann & Schallberger 2000b[71])

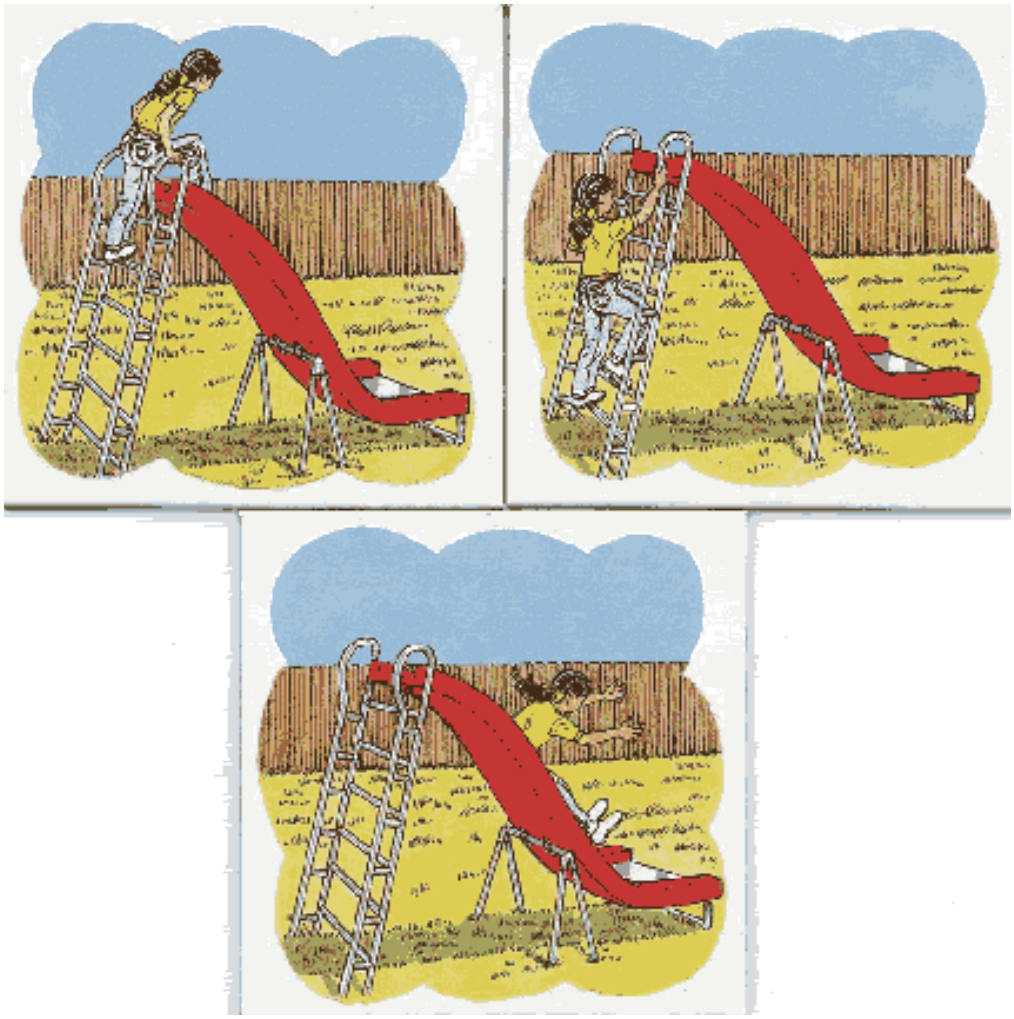

Abbildung 3.4: Beispiel aus dem Untertest Bilderordnen (aus: Tewes, Rossmann & Schallberger 2000b)

[71] Die nachfolgenden drei Beispielaufgaben stammen aus dem HAWIK-III Testkoffer (vgl. Tewes, Rossmann & Schallberger 2000b).

Figurenlegen (FL): Die Schülerin oder der Schüler soll nacheinander zehn in verschiedene Teile zerschnittene Figuren richtig zusammensetzen. Dieser Untertest prüft die visuell-motorische Koordination und die Fähigkeit, Beziehungen zu erfassen.

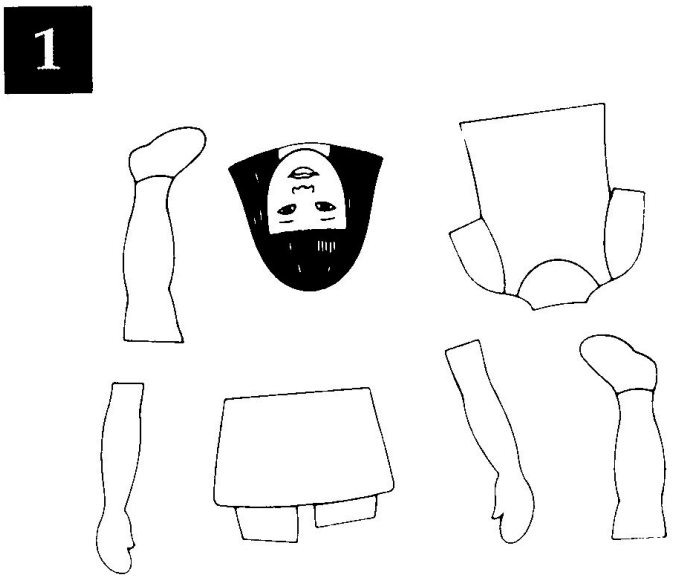

Abbildung 3.5: Beispiel aus dem Untertest Figurenlegen (aus: Tewes, Rossmann & Schallberger 2000b)

Zahlen-Symbol- Test (ZS): Einem bestimmten Schlüssel entsprechend sollen die Lernenden den Zahlen 1-9 jeweils ein bestimmtes Symbol zuordnen. Dieser Untertest gibt Auskunft über die Arbeitsgeschwindigkeit, die visuell-motorische Koordination, die Lernfähigkeit und das Konzentrationsvermögen.

Symbolsuche (SS): Innerhalb von 2 Minuten sollen durch Vergleiche möglichst viele Symbolsequenzen auf das Vorhandensein je eines vorgegebenen Symbols überprüft werden. Dieser Untertest ist ein Indiz für die Arbeitsgeschwindigkeit.

Labyrinth-Test (LT): Die Schülerinnen und Schüler müssen in ein auf Papier vorgegebenes Labyrinth mit Bleistift den Weg vom Start am Rande in das Ziel im Inneren aufzeichnen. Obwohl die Kinder diesen Test sehr schätzen, kommt er selten zum Zug. Er wird lediglich dann verwendet, wenn ein regulärer Untertest des Handlungsteils nicht durchgeführt oder ausgewertet werden konnte. Deswegen wurde er in der Abbildung auch nicht erwähnt (vgl. Abbildung 3.2).

3.3.3 Interpretation der Testergebnisse

Die Interpretation der Ergebnisse von beiden Untertestteilen erfolgt getrennt, wobei ein *Verbal-IQ* sowie ein *Handlungs-IQ* berechnet werden. Das Ergebnis des Verbalteils bezeichnet man häufig als *Sprachliche Intelligenz*, das Ergebnis des Handlungsteils als *Praktische Intelligenz*.

- *Sprachliche Intelligenz* beinhaltet sprachliche Konzeptbildung, sprachliche Kenntnisse und verbales Ausdrucksvermögen (z.B. Beantwortung von mündlich gestellten Fragen zur Prüfung des Faktenwissens, von Wortbedeutungen, des schlussfolgernden Denkens und der Fähigkeit, die eigenen Gedanken in Worte zu fassen);
- *Praktische Intelligenz* beinhaltet Wahrnehmungsorganisation (z.B. nonverbale Denkvorgänge und visuell-motorische Koordination) sowie Arbeitsgeschwindigkeit (z.B. Denkgeschwindigkeit

und motorische Geschwindigkeit bei der Bearbeitung verschiedener nonverbaler Aufgaben mit geringer Schwierigkeitskomponente).

Diese Unterscheidung zwischen sprachlichen und handlungsbezogenen Leistungsaspekten ist sinnvoll, da auf diese Weise geprüft werden kann, ob der Gesamt-IQ in erster Linie auf gute sprachliche Begabungen oder eher auf gute praktische Begabungen zurückzuführen ist, oder ob beide Aspekte etwa gleich stark ausgeprägt sind.

Wenn die oder der Lernende in nahezu allen Leistungsbereichen weit überdurchschnittliche Leistungen zeigt, erzielt sie, er einen sehr hohen *Gesamt-IQ*. Ebenso wird ein sehr niedriger Gesamt-IQ nur dann erzielt, wenn die Schülerin oder der Schüler auf sehr niedrige Leistungen in fast allen Bereichen kommt.

Wenn der IQ im mittleren Bereich liegt, d.h. sich einem Wert von 100 Punkten nähert, kann die oder der Lernende ein sehr homogenes oder ein sehr heterogenes Leistungsbild aufweisen. Sie oder er kann durchschnittliche Leistungen in allen Untertests erzielt haben, oder weit unterdurchschnittliche als auch weit überdurchschnittliche Leistungen in einzelnen Teilleistungsbereichen. Um ein differenziertes Leistungsbild einer solchen Lernerin, eines solchen Lerners machen zu können, ist die Einsicht in die Ergebnisse der einzelnen Untertests nötig.

Somit sollte bei der Interpretation eines Gesamt-IQ-Wertes immer mitberücksichtigt werden, wie dieses Resultat zustande gekommen ist.

3.3.4 Gütekriterien des Intelligenztestes HAWIK-III

Objektivität
Die Objektivität ist gewährleistet, wenn sich die Untersuchungsleitenden an das Handbuch und die darin vorgeschriebenen Kriterien halten. Das bedeutet, dass die Person, welche die Messungen vornimmt, kaum einen Einfluss auf das Testresultat hat, wenn sie den Instruktionen aus dem Handbuch strikt folgt.

Reliabilität
Die interne Konsistenz der Untertests variiert in ihrer Zuverlässigkeit zwischen r = .68 und r = .88. Die Reliabilität des Gesamttests beträgt r = .96 und ist somit sehr hoch, d.h. es ist ein ausserordentlich guter Test. Der Messfehler des IQ entspricht bei einer 95-prozentigen Wahrscheinlichkeit einem Fehlerbereich von +/- 6.20, was nach den allgemeinen Anforderungen an Tests recht hoch ist.

Validität
Die Kriteriumsvalidität, d.h. die Übereinstimmung mit den Ergebnissen aus anderen Datenquellen (z.B. Tests, Noten), ist gegeben. Die Korrelationen zwischen IQ-Werte, die mittels HAWIK erhoben wurden und Noten und Lehrerinnen- und Lehrerurteil sind mittelhoch, wobei der Gesamt-IQ den besten Prädikator darstellt.

3.4 Anwendungsmöglichkeiten: Intelligenztests bei Lernstörungen

Obwohl es aufgrund der bisherigen Darstellungen so erscheinen mag, können Intelligenztests keinesfalls nur im Falle von (vermuteter) Hochbegabung eingesetzt werden.

Betrachten wir die Normalverteilungskurve des Merkmals Intelligenz: Der Mittelwert beträgt 100 Intelligenzpunkte, um die herum sich 68% der Bevölkerung streuen. Das heisst zwei Drittel aller Menschen erreichen einen Intelligenzquotienten zwischen 85 und 115. Einen IQ zwischen 70 und 130 haben etwa 95% der Bevölkerung, die restlichen 5% befinden sich zur Hälfte drunter bzw. drüber.

Wird nun bei einer Schülerin oder einem Schüler eine Lernstörung vermutet, kann anhand eines Intelligenztests diagnostiziert werden, ob sie oder er auch einen ausserordentlich tiefen IQ und somit eine *Minderbegabung* aufweist. Hier muss noch angefügt werden, dass der diagnostizierte Intelligenzquotient nicht das alles entscheidende Kriterium sein darf in der darauf folgenden Entscheidungsfindung. Weisen die getesteten Schülerinnen und Schüler trotz schulischer Schwächen eine Hochbegabung auf, sollen sie gezielt gefördert werden. Kann die Lernstörung auf einen oder mehrere bestimmte Bereiche zurückgeführt werden, handelt es sich um eine *Teilleistungsschwäche* bei einer durchschnittlichen Intelligenz, auf welche mit Zusatzunterricht oder anderer spezifischer Förderung Rücksicht genommen werden sollte.

Im Folgenden wird die Thematik der Teilleistungsschwäche anhand zweier Beispiele dargestellt.

3.4.1 Legasthenie

Bereits anfangs des 20. Jahrhunderts befassten sich Ärzte mit jenen Kindern und Jugendlichen, die nicht lernten, richtig zu lesen und nach Diktat zu schreiben, obwohl sie sonst relativ unauffällig waren. Solche *Lese- und Rechtschreibschwierigkeiten* nennt man *Legasthenie*; die Betroffenen sind Legastheniker oder Legasthenikerinnen. Versuche, eine einheitliche Symptomatik der Legasthenie herauszufiltern, sind gescheitert (vgl. Ingenkamp 1997, p. 164). Da es also keinen Standard und keine Norm gibt, die für Legasthenie gültig sind, könnte man formal sagen, Legasthenie gibt es nicht. Natürlich gibt es sie doch, nur existiert sie als persönliche Beschaffenheit, sie ist ein individuelles Phänomen (vgl. Steltzer 1998). Inzwischen wird auch darauf verzichtet, einen Zusammenhang zur Intelligenz feststellen zu wollen, denn Legasthenie wird auf den unterschiedlichsten Intelligenzniveaus beobachtet.

> «Unter Legasthenie verstehen wir eine spezielle, aus dem Rahmen der übrigen Leistungen fallende Schwäche im Erlernen des Lesens (und indirekt auch des selbständigen fehlerfreien Schreibens) bei sonst intakter oder – im Verhältnis zur Lesefähigkeit – relativ guter Intelligenz» (Linder 1951; zit. nach Ingenkamp 1997, p. 164).

Die Schwierigkeit liegt nun einerseits darin, die diversen Ursachenschemata mit Berücksichtigung der unterschiedlichen Gewichtungen einzelner Ursachen in diagnostische Strategien umzusetzen, andererseits ist diese Breite des diagnostischen Ansatzes in der Schulpraxis nicht zu realisieren. Es bestand daher das Bedürfnis, legastheniespezifische Fehler im Lesen und Schreiben als Diagnosekriterien einzusetzen (vgl. Ingenkamp 1997, p. 167ff.). Dies konnte jedoch nicht beibehalten werden, da sich in Folgeuntersuchungen keine Fehlerarten nachweisen liessen, die nicht auch bei Kontrollstichprobe (oder umgekehrt) in Erscheinung treten. Zudem wird gegenwärtig davon ausgegangen, dass Lese- und Rechtschreibschwierigkeiten keineswegs immer gekoppelt sein müssen; Häufigkeit und Resistenz der Fehler sind daher massgeblich (vgl. ebd., p. 166).

Es hat sich also gezeigt, dass bei einer so uneinheitlichen Teilleistungsschwäche neben dem Lehrerinnen- und Lehrerurteil andere diagnostische Hilfsmittel unabdingbar sind. Neben verschiedenen formellen Tests[72] wird aus diesem Grunde mit der Intelligenzdiagnostik gearbeitet. Das Ergebnis des Intelligenztests kann wichtig für die *Art der Förderung* sein. Bei einem IQ-Wert unter 85 muss sorgfältig überprüft werden, ob eine *Sonderschulbedürftigkeit* vorliegt. Ist dem nicht so und/oder liegt der IQ im Bereich von ungefähr 85-100, wird man sich auf ein anderes Lehren bzw. Lernen einstellen müssen. Bei Verdacht auf Legasthenie sollte auf jeden Fall ein Intelligenztest mit nonverbaler Aufgabenstellung eingesetzt werden (vgl. ebd., p. 169).[73]

---

[72] Eine ausführliche Auflistung ist in Ingenkamp (1997, p. 167ff.) zu finden.

[73] Ingenkamp (1997, p. 169f.) schlägt Gruppentests, Bildertests und standardisierte Lautbildungs- und Lautunterscheidungstests vor.

3.4.2 Dyskalkulie

Nebst den Lese- und Rechtschreibschwierigkeiten (LRS) ist häufig auch die *Rechenschwäche Dyskalkulie*[74] anzutreffen.

> «Dyskalkulie ist definiert als eine spezielle Rechenstörung, die sich vom Rechenversagen bei einer allgemeinen Schulleistungsschwäche abhebt. Dyskalkulie wird eine Rechenschwäche genannt, wenn ein Kind in der Schule vorwiegend im Rechnen Schwierigkeiten hat und in den übrigen Fächern gute bis sehr gute Leistungen erbringen kann» (Meile & Lüchinger 1983; zit. nach Schilling & Prochinig 1998).

Zur Abklärung dieser Teilleistungsschwäche wird ebenfalls eine Intelligenzmessung durchgeführt. Aus der Kollektivuntersuchung von Grissemann und Weber (1982) hat sich ergeben, dass 9,5% der Lernenden bei einem Gesamtanteil von 20,4% rechenschwacher Schülerinnen und Schüler einen IQ im Bereich von 90-99 haben. Bei 7,5% liegt er unter 90, 4% liegen im IQ-Bereich von 100 und mehr. Dyskalkulie kann eigentlich erst in der Schule, das heisst in den ersten beiden Schuljahren und später festgestellt werden. Es ist überaus wichtig, dass ein rechenschwaches Kind so früh wie möglich erkannt wird und Massnahmen ins Auge gefasst werden, um diese Teilleistungsschwäche zu beheben oder zumindest zu lindern (vgl. Schilling & Prochinig 1998).

3.5 Kritik am Einsatz von Intelligenztests

Am Beginn der modernen Intelligenzdiagnostik stand eine schulpraktische Aufgabe: die Auswahl sonderschulbedürftiger Schülerinnen und Schüler. Der von Binet und Simon für diesen Zweck entwickelte Intelligenztest ist das prominenteste Diagnoseinstrument ganzer Testgenerationen geworden. Mit ihm hat sich die Frage vererbt, ob eine schulische Auslese mit Hilfe von Intelligenztests statthaft sei. Dem heutigen Wissensstand bei der *Schuleingangsdiagnostik* (Abklärung der «Schulreife») entsprechend kann man antworten, dass ein Intelligenztest alleine nicht ausreicht (vgl. Ingenkamp 1997, p. 195f.).

3.5.1 Kritik am Einsatz von Intelligenztests im Zusammenhang mit Hochbegabung

Weil sich Hochbegabung nicht ausschliesslich über hohe Intelligenz definieren lässt, ist der Intelligenztest als einziges Identifikationsverfahren abzulehnen. In Anlehnung an Ey-Ehlers (2001, p. 84f.), die sich auf diverse andere Autoren und Autorinnen bezieht, werden im Folgenden einige Mängel des Intelligenztests aufgezeigt:

- Es wird einzig die kognitive Komponente und ein logisch-analytisches, konvergentes Denken gemessen, andere Determinanten der Hochbegabung – wie zum Beispiel divergentes Denken – werden nicht einbezogen (vgl. ebd.).
- In Tests erreichte und in alltäglichen Lebenssituationen gezeigte Intelligenzleistungen liefern unterschiedliche Beurteilungskriterien bei Hochbegabung (vgl. ebd., p. 85)
- Hochbegabte zeichnen sich in ihrem Denken nicht nur durch korrekte Lösungen aus, sondern auch durch *niveauvollere Lösungsstrategien*. Dieser Idee zugrunde liegend versucht eine Forschungsrichtung innerhalb der Begabungsdiagnostik, über die reine Erhebung der «Denkprodukte» hinauszugehen. Die Ergebnisse von Denkprozessen, die durch richtige Antworten in Intelligenztests zum Ausdruck kommen, stellen also nicht mehr das anzustrebende Ziel dar. Zudem sind die in Tests geforderten Denkprozesse meist sehr eingegrenzt und bilden nicht die Denkprozesse ab, die bei real auftretenden Problemen zu guten Lösungen führen. Intelligentes

[74] Neben den verwendeten Ausdrücken Dyskalkulie und Rechenschwäche sind in der Literatur die gleichbedeutenden Begriffe Akalkulie, Anarithmie, Rechenlegasthenie und Arithmasthenie anzutreffen.

Verhalten zeigt sich in der Praxis oft in der Bewältigung strukturierter Probleme, bei denen es keine definierte «richtige» Lösung gibt. Weil somit die Breite und Vielfalt intellektueller Begabung mit den herkömmlichen Verfahren zur Intelligenzmessung nicht entdeckt werden kann, müssen andere Ansätze zum Zug kommen (vgl. Holling & Kanning 1999, p. 52).[75]

3.5.2 Kritik an den Intelligenztests

Sieht man vom Einsatz des Intelligenztests bei Hochbegabung einmal ab, können noch weitere grundlegende Mängel aufgedeckt werden (vgl. Ey-Ehlers 2001, p. 84f.):

- Die Ergebnisse der Intelligenztests berücksichtigen nie die *Entwicklungsmöglichkeiten* eines Individuums, sondern nur den gegenwärtigen Status der getesteten Person. Da Intelligenz eine durch Lern- und Entwicklungseinflüsse veränderbare Grösse ist, sollen zur Bestimmung der intellektuellen Leistung nicht nur die gelösten Aufgaben berücksichtigt werden, «sondern auch die, die nach *zusätzlichen Hilfestellungen* bewältigt werden» (Kormann 1982; zit. nach Lukesch 1998, p. 260) können. Solche so genannten *Lerntests* werden zur Verbesserung der konventionellen Intelligenztests eingesetzt und versuchen die intellektuelle Lernfähigkeit zu erfassen. Der grundsätzliche Unterschied zu Intelligenztests besteht in der Struktur des Lerntests: nach einem Vortest erfolgt eine «Pädagogisierungsphase» (Lukesch 1998, p. 260), in der die Aufgabenlösung besprochen wird. Erst die Resultate des anschliessenden Nachtests sind letztendlich auf ihren prognostischen Wert zu prüfen. Auf diese Weise sollten Unterschiede in Bezug auf die Erfahrung mit Tests ausgeglichen werden. Zudem resultiert durch die mehrmalige Leistungserhebung ein reliableres Mass der Lern- und Leistungsfähigkeit eines Individuums als bei einer punktuellen Erhebung (vgl. ebd., p. 260f.).
- Die *Kultur- und Schichtspezifität der Tests* ist auf die Mittel- und Oberschicht ausgerichtet und benachteiligt Schülerinnen und Schüler aus der Unterschicht bzw. eine Reihe von Minoritäten (vgl. Heinbokel 1988).
- Lernende mit ungeübten oder unzureichenden Sprachfertigkeiten sind in verbalen Intelligenztests benachteiligt (vgl. Lukesch 1995). Selbst wenn der Test gar keine verbalen Anteile enthält, kann es sein, dass diese Lernenden die Anweisungen nicht verstehen (vgl. Rost 1989).
- Die getesteten Kinder können unter Testangst oder momentanen stimmungsmässigen und gesundheitlichen Beeinträchtigungen leiden, ebenso können die durch die Testsituation entstehenden Störfaktoren Konzentration und Aufmerksamkeit beeinträchtigen (vgl. Feger 1988).
- Ein weiterer Kritikpunkt ist die mögliche Beeinträchtigung der Messergebnisse, da durch die Addition der Untertestwerte zu einem einzigen Punktwert Teilleistungsstärken bzw. -schwächen verdeckt werden (vgl. Winner 1998).

---

[75] Zu nennen sind u.a. computergestützte Planspiele, wo die Spielerinnen und Spieler mit einem komplexen System zurechtkommen müssen, indem sie bei ihren Entscheidungen die Auswirkungen auf das gesamte System mit berücksichtigen. Zum Beispiel sollen die Spielerinnen und Spieler sich bei einem simulierten Wirtschaftsunternehmen in die Rolle einer Betriebsleiterin oder eines Betriebsleiters versetzen und den Betrieb über 15 simulierte Monate leiten. Während des Planspiels sollen sie ihre Überlegungen verbalisieren. Auf diese Weise können ihre Analysen, Ziele und Pläne erfasst werden (vgl. Holling & Kanning 1999, p. 52f.).

3.6 Fazit

Es soll hier nicht der Anschein erweckt werden, dass eine grundsätzliche Aversion gegen Intelligenztests besteht. Dennoch soll die Kritik am Einsatz von Intelligenztests aufzeigen, dass IQ-Zahlen und Intelligenzprofile nicht als endgültig zu werten sind. Sie sind stark abhängig von der allgemeinen psychischen und der sozio-kulturellen Situation der getesteten Person sowie von der spezifischen Testsituation (vgl. Jegge 1976, p. 44).

Die Darstellung der unterschiedlichen Verfahren zur Hochbegabungsdiagnostik veranschaulicht, dass *ein einziges Verfahren die vielfältigen Erscheinungsformen der Hochbegabung nicht erfassen kann*. Vielmehr lassen sich durch jedes Verfahren Informationen gewinnen, die anhand anderer Instrumente nicht erhoben werden können. Vor diesem Hintergrund stellt eine Diagnostik, die *multimodal* und *multimethodal* ausgelegt ist, eine optimale Lösung dar (vgl. Holling & Kanning 1999, p. 53).

Es scheint angebracht, an dieser Stelle mit einem Zitat von Kastner (1982) abzuschliessen:

> «Festzuhalten bleibt für die Pädagogische Diagnostik, dass ‹Intelligenz› einen schillernden Begriff darstellt, der in keiner Weise der quantitativen und qualitativen Vielfalt von Einzelfähigkeiten und deren Entwicklung in der Interaktion zwischen Individuum und Umwelt gerecht wird» (ebd., p. 338).

4 Unterrichtsbeobachtung

4.1 Fallbeispiel Mobbing in der Klasse

Beat Anderegg[76] ist seit wenigen Monaten Klassenlehrer an einer fünften Klasse in einem Dorfschulhaus in ländlicher Umgebung. Unmittelbar vor seiner Anstellung hat er seine Ausbildung abgeschlossen. Den lange ersehnten Einstieg in den Beruf hat Anderegg sehr gut erlebt. Zu den Schülerinnen und Schülern scheint er den Draht gefunden zu haben, eine gute Basis für Elternkontakte und die Zusammenarbeit im Lehrerkollegium ist gelegt.

Seit einigen Wochen fällt dem Lehrer jedoch auf, dass sich Kevin nur noch selten am mündlichen Unterricht beteiligt und oft seine Unterrichtsmaterialien sucht. In einem Gespräch verspricht Kevin, in Zukunft sein Pult besser in Ordnung zu halten und sich vermehrt am Unterricht zu beteiligen. Während des Gesprächs macht Kevin einen verängstigten und eingeschüchterten Eindruck. In den folgenden Wochen erkennt der Lehrer die Bemühungen Kevins, die Situation ändert sich jedoch nicht grundsätzlich. Herr Anderegg thematisiert Kevins Verhalten in einer Pause im Lehrerzimmer und erhofft sich, dadurch mehr Informationen über Kevins Schulverhalten zu erlangen. Er kann in Erfahrung bringen, dass Kevins Eltern erst vor einem Jahr zugezogen sind und dass Kevin anfangs Mühe hatte, sich in die neue Klasse zu integrieren. Die Lehrerin der vierten Klasse kann sich an Situationen erinnern, in welchen Kevin von zwei seiner Mitschüler, Joël und Luca, mehrmals geärgert wurde. Anderegg entschliesst sich, das Verhalten von Joël und Luca gegenüber Kevin genauer zu beobachten. In der Hektik des Unterrichts beschränken sich seine Beobachtungen aber häufig nur auf die Momente, in welchen Kevin seine Unterrichtsmaterialien nicht zur Hand hat. Analysiert er diese Situationen, erkennt er jedoch keinen sichtbaren Zusammenhang mit Verhaltensweisen anderer Schülerinnen und Schüler. Tage später entdeckt der Lehrer beim Entsorgen des Altpapiers mehrere Arbeitsblätter und ein Heft von Kevin im Altpapier. Eine spätere Umstellung der Zimmereinrichtung fördert überdies den vermissten Hausschuh Kevins zutage. Der Lehrer wird sich bewusst, dass Kevin nicht alleine für das Fehlen seiner Unterrichtsmaterialien verantwortlich sein kann. Die Reflexion der Situation bringt Anderegg auf den Gedanken des Mobbings[77]. Beim Sichten seiner Studienunterlagen zum Thema Mobbing wird ihm erneut bewusst, wie viele Gesichter Mobbing haben kann und welche Grundzüge ihm eigen sind. Zu diesen Grundzügen gehören die Erniedrigung der Opfer, das Schweigen der Opfer aufgrund eines Schamgefühls oder aus Angst vor weiteren Repressionen, sowie die Hilflosigkeit und Isolation der Opfer (vgl. Alsaker 2003, p. 24ff.).

Einige der Merkmale, die für Mobbing sprechen, treffen auf Kevin zu. Viele Episoden scheinen sich jedoch sehr subtil und nicht sichtbar abzuspielen und bleiben dadurch für den Gelegenheitsbeobachter unerkannt. Nach diesen ersten Erkenntnissen entschliesst sich der Lehrer, das Unterrichtsgeschehen systematischer und gezielter zu beobachten und das Gesehene protokollarisch festzuhalten. Neben Beobachtungen des Unterrichts im Klassenzimmer weitet er seine Aufmerksamkeit auf den Sportunterricht und die Pausenaktivität der Kinder aus. Schon nach kurzer Beobachtungsdauer fällt ihm auf, dass Kevin auf dem Pausenplatz und im Sport von einer kleinen Gruppe von Knaben aus der Klasse gemieden wird. Beim Zusammenstellen der Spielmannschaften achten diese Knaben darauf, dass Kevin nicht in ihre Mannschaft gewählt wird. Ist er trotzdem Teil ihrer Mannschaft, so wird er für begangene Spielfehler unverhältnismässig oft getadelt und nach Möglichkeit vom Spielgeschehen ausgegrenzt. Es ist dieselbe Knabengruppe, die Anderegg in den kleinen Pausen ab und zu mit zusammengesteckten Köpfen im Flur entdeckt. Interessanterweise

---

[76] Alle in diesem Fallbeispiel verwendeten Namen sind frei erfunden.

[77] Von Mobbing (dt.: Plagen) wird gesprochen, «wenn ein Kind wiederholt und systematisch den direkten oder indirekten negativen Handlungen eines oder mehrerer Kinder ausgesetzt ist» (Alsaker 2003, p. 19).

häufen sich danach die Vorfälle, in welchen Kevin Unterrichtsmaterial oder persönliche Gegenstände vermisst.

Mit den angefertigten Beobachtungsprotokollen sucht Anderegg Hilfe beim schulpsychologischen Dienst. Die protokollierten Beobachtungen ermöglichen einen raschen und guten Einblick in die Klassensituation und sind dadurch für die weitere Problemlösung von grossem Nutzen.

Dieses Fallbeispiel soll im Folgenden dazu dienen, das Vorgehen, die Methodik und Möglichkeiten sowie auch die Grenzen der Unterrichtsbeobachtung zu verdeutlichen.

4.2 Verhaltensbeobachtung

Die Verhaltensbeobachtung stellt ein zentrales Element der pädagogischen Diagnostik dar. Eine Vielzahl wichtiger Informationen über das Unterrichtsgeschehen gewinnen Lehrerinnen und Lehrer aufgrund zahlreicher Verhaltensbeobachtungen.

Nach einer Definition der Begriffe Beobachtung und Wahrnehmung (Abschnitt 4.2.1) sollen in Abschnitt 4.2.2 die Unterschiede zwischen systematisch und unsystematisch angelegten Beobachtungen dargestellt werden. Abschliessend werden methodische Aspekte thematisiert, die bei der Planung und Durchführung von Verhaltensbeobachtungen berücksichtigt werden müssen (Abschnitt 4.2.3). Die Ausführungen dazu beschränken sich auf den Beobachtungsgegenstand, das Beobachtungsfeld und die Beobachtungseinheit.

4.2.1 Beobachtung

Unser alltägliches Leben ist geprägt von aktiven, nicht zielgerichteten Wahrnehmungsprozessen. Sie informieren uns über die Abläufe in der uns umgebenden Realität. Dabei können bereits gemachte Erfahrungen und Erkenntnisse die Selektion und Entschlüsselung des Wahrgenommenen so beeinflussen, dass Menschen, Gegenstände und Situationen eine jeweils individuelle Bedeutung für den Betrachter erhalten. Wahrnehmung wird dadurch ein von Subjektivität geprägter Selektionsprozess (vgl. Martin & Wawrinowski 1991, p. 13ff.).

Wird Wahrnehmung zu einer zielgerichteten und beabsichtigten Handlung, kann von Beobachtung gesprochen werden. Diese ist motiviert durch persönlichkeitsspezifische Erwartungen, Interessen, Erfahrungen und Vorurteile. Beobachtung ist dementsprechend kein passives Auf-sich-einwirken-Lassen von Sinnesreizen, sondern eine Form des aktiven Sich-Aneignens von Wirklichkeit, mit der Absicht, dadurch Informationen über den Beobachtungsgegenstand und sein Umfeld zu erhalten (ebd., p. 32f.).

Die Verhaltensbeobachtung ist nach Roth «das grundlegende Verfahren der empirisch forschenden Sozialwissenschaften und der Erziehungswissenschaften» (Roth 1978, p. 80). Durch die Tätigkeit des Beobachtens lassen sich Erkenntnisse über Verhaltensweisen von Menschen in Erfahrung bringen, die durch eine Befragung nur in Form eines Reports erhoben werden können, welche keineswegs mit dem effektiven Verhalten der befragten Person übereinstimmen muss (vgl. Lamnek 2001, p. 266). Insofern liefern Beobachtungen tendenziell validere Informationen zu *Verhaltensweisen* als Befragungen. Andererseits lassen sich *Verhaltensdispositionen*, wie beispielsweise Motive, Einstellungen und Wertungen eines Menschen nicht unmittelbar beobachten. Allenfalls lassen sich durch die Handlungen des Beobachteten Rückschlüsse auf seine Befindlichkeit machen. Solche Rückschlüsse verfügen jedoch über einen hohen Anteil an Spekulativität (vgl. Ledl 1997, p. 13ff.).

Verhaltensbeobachtungen werden aber nicht nur im wissenschaftlichen Kontext angewandt, sondern sind integrierter Bestandteil der Unterrichtsbeobachtung und dadurch von besonderer Bedeutung für die pädagogische Diagnostik. Die Aufgaben der Institution Schule und die daraus

resultierenden Ansprüche im Bereich der Wissensvermittlung, der Selektion und der Sozialisation verlangen von der Lehrperson regelmässige Verhaltensbeobachtungen der Schülerinnen und Schüler. Aufgrund dieser Beobachtungen lassen sich die für die Beurteilung und Förderung wichtigen Informationen betreffend Entwicklungsstand und individuellem Fortschritt der Auszubildenden gewinnen (vgl. Ingenkamp 1997, p. 53). Wie im einleitenden Fallbeispiel gezeigt, können sich im Schulalltag zusätzliche und spezifischere Beobachtungsbedürfnisse herauskristallisieren. Häufig entstehen solche Verhaltensbeobachtungen aufgrund zufällig gemachter Wahrnehmungen, die ein Interesse zur genaueren Überprüfung mittels gezielter Beobachtung wecken (vgl. Martin & Wawrinowski 1991, p. 29).

Die Beobachtenden sind bei der Datenerhebung das wesentliche Messinstrument und können durch ihre subjektiv geprägte Selektions- und Interpretationsweise eine Fehlerquelle darstellen. Unsystematische und zufällige Beobachtungen erhöhen die Gefahr einer subjektiven Wahrnehmung und Interpretation der beobachteten Situation, während ein systematisches und regelgeleitetes Beobachten die Subjektivität zu begrenzen und zu kontrollieren versucht (vgl. von Sydow 2004, p. 8). Der folgende Abschnitt thematisiert die Unterschiede zwischen unsystematischem und systematischem Beobachten und zeigt Kriterien für ein wissenschaftliches Beobachten auf, welches die Forderungen der Validität, Reliabilität und Objektivität berücksichtigt.

4.2.2 Unterschiede zwischen unsystematischer und systematischer Beobachtung

In unserem Fallbeispiel nimmt der Klassenlehrer ohne bestimmte Beobachtungsabsichten eine Fülle von Ereignissen wahr. Viele dieser Wahrnehmungen finden in Form eines «Zuschauens» statt und sind daher nicht systematisch. Solche Beobachtungen werden in der Fachliteratur als naive oder auch unsystematische Beobachtungen (synonyme Begriffe: Gelegenheitsbeobachtung, Alltagsbeobachtung) umschrieben. Sie sind ungerichtet und verfolgen keine bestimmte Zielsetzung (vgl. Ingenkamp 1997, p. 53). Sie werden beiläufig gemacht und verfügen in den meisten Fällen über wenig wichtigen Gehalt. Sie können jedoch von Interesse werden, wenn sich ein zufällig beobachtetes Schülerinnen- oder Schülerverhalten (bspw. nonkonformes Verhalten) zu häufen beginnt und die Lehrperson aufgrund dieser Entwicklung ihre Beobachtung zu systematisieren beginnt. So misst der Klassenlehrer aus dem Fallbeispiel den ersten, zufällig gemachten Beobachtungen von Kevin keine besondere Bedeutung bei. Er hat Kevin bisher als aufmerksamen und zuverlässigen Schüler erlebt. Erst eine Häufung der Fälle und das Auffinden einiger Unterrichtsmaterialien im Altpapier lassen bei der Lehrperson den Entschluss reifen, Kevin und auch die sozialen Interaktionen innerhalb der Klasse systematischer zu beobachten (vgl. Martin & Wawrinowski 1991, p. 37).

Die systematische respektive wissenschaftliche Beobachtung verfolgt ein klar umrissenes und theoretisch begründetes Untersuchungsziel. Wissenschaftliche Beobachtungen müssen dem Nachweis, der Bestätigung oder der Widerlegung von Hypothesen dienen (ebd., p. 37f.). Eine Beobachtung kann nach Atteslander (1995, p. 126) nur dann wissenschaftlich sein, wenn sie einem bestimmten Forschungszweck dient und den theoriebestimmten Zugang zur Forschungsfrage ermöglicht. Dabei wird die Beobachtung durch die Festlegung der Beobachtungszeiten, der Beobachtungseinheiten und des Grades der Teilnahme des Beobachters systematisch geplant. Eine systematische Aufzeichnung vorher bestimmter Ereignisse soll eine Kontrollierbarkeit schaffen und Wiederholbarkeit der Beobachtung ermöglichen. Als weiteres Kriterium müssen wissenschaftliche Beobachtungen wiederholten Prüfungen und Kontrollen hinsichtlich der Objektivität, Validität und Reliabilität standhalten können. Diese Gütekriterien sind nur gegeben, «wenn mehrere Beobachter bei gleichzeitiger Beobachtung desselben Beobachtungsgegenstandes zu übereinstimmenden Beurteilungen gelangen» (Köhne 1979, p. 24). Ergänzt werden könnte diese Definition durch den Anspruch einer zielgerichteten und selektiven Vorgehensweise, welche einen konkreten Beobachtungsgegenstand

fokussiert und andere Aspekte bewusst vernachlässigt (vgl. Lamnek 2001, p. 265). Wissenschaftliche und systematische Beobachtungen lassen sich damit als «*das systematische Erfassen, Festhalten und Deuten sinnlich wahrnehmbaren Verhaltens zum Zeitpunkt seines Geschehens*» (Atteslan- (Atteslander 1995, p. 87; Hervorhebungen im Original) verstehen.

4.2.3 Methodische Aspekte des Beobachtens

Die wissenschaftliche und systematische Beobachtung beinhaltet nach Lamnek verschiedene methodische Aspekte, die bei der Planung und Durchführung berücksichtigt werden müssen (2001, p. 270ff.). Es sind dies der Beobachtungsgegenstand, das Beobachtungsfeld und die Beobachtungseinheit. Diese methodischen Aspekte sind jedoch nicht nur für wissenschaftlich durchgeführte Beobachtungen relevant, sondern haben auch eine wesentliche Bedeutung für die Unterrichtsbeobachtung einer Lehrperson.

Vor einer Beobachtungssequenz stellt sich die Frage nach dem *Beobachtungsgegenstand*. Wie bereits in Abschnitt 4.2.2 erwähnt, zielt die Beobachtung darauf ab, Verhalten zum Zeitpunkt seines Geschehens festzuhalten. Je komplexer das zu beobachtende Verhalten konstituiert ist, desto wichtiger wird eine vertiefte inhaltliche Auseinandersetzung mit dem Beobachtungsgegenstand. Eine Unterrichtsbeobachtung zu Mobbing beispielsweise bedarf einer eingehenden Auseinandersetzung mit entsprechender Fachliteratur. Erst die fundierten und theoriegeleiteten Kenntnisse über einen Beobachtungsgegenstand ermöglichen eine sinnvolle Planung und Durchführung einer Beobachtungseinheit (vgl. Ingenkamp 1997, p. 56ff.).

Als *Beobachtungsfeld* bezeichnet Atteslander (vgl. 1995, p. 97) denjenigen räumlichen und/oder auch sozialen Bereich, in welchem die Beobachtungen stattfinden sollen. Das Beobachtungsfeld definiert sich jedoch nicht ausschliesslich über den räumlichen Aspekt einer Beobachtung, sondern auch über die Frage, wann und unter welchen Rahmenbedingungen beobachtet werden soll (vgl. Lamnek 2001, p. 271). Das Beobachtungsfeld einer Lehrperson entspricht einer Feldbeobachtung (vgl. Abschnitt 4.3.5) und ist entsprechend an die natürlichen Gegebenheiten des Unterrichts und die dort herrschenden räumlichen, zeitlichen und situativen Bedingungen gebunden. Dabei ist in Rechnung zu stellen, dass die hohe Komplexität des Umfeldes die Lehrperson dazu zwingt, sich in ihren Beobachtungen einzuschränken.

Diese Einschränkungen führen zur Festlegung von *Beobachtungseinheiten*, die klären, wer oder was zu welcher Zeit beobachtet werden soll. Nach Atteslander (1995, p. 99) bezeichnet die Beobachtungseinheit «denjenigen Teilbereich sozialen Geschehens, der konkreter Gegenstand der Beobachtung sein soll». Die Dauer der Beobachtungseinheiten ergibt sich aufgrund der zu erwartenden Auftretenshäufigkeit der interessierenden Verhaltensmerkmale, der Genauigkeit der Aussagen, die erreicht werden soll und den zur Verfügung stehenden Ressourcen. Gerade für die Unterrichtsbeobachtung ist es aufgrund der Komplexität, der Unmittelbarkeit von Unterricht und den beschränkten Ressourcen[78] wichtig, möglichst kleine Beobachtungseinheiten festzulegen.

4.3 Formen der Beobachtung

Eine in sich geschlossene Theorie wissenschaftlicher Beobachtungen existiert bislang nicht. Methodische Ausführungen zur Beobachtung als Forschungsmethode beschränken sich oft auf eine Klassifikation (taxonomische Darstellung) verschiedener Beobachtungsformen (vgl. Lamnek 2001, p. 267ff.). In der Folge sollen einige dieser Beobachtungsformen dargestellt werden.

---

[78] Die Lehrperson ist Unterrichtende und Beobachtende zugleich (vgl. Abschnitt 4.3.1).

4.3.1 Teilnehmende vs. nicht-teilnehmende Beobachtung

Ein Kriterium zur Klassifikation systematischer Beobachtungen stellt die Frage nach der Distanz des Beobachters zum Geschehen dar (vgl. Kleber 1992, p. 199). Von einer *teilnehmenden Beobachtung* wird gesprochen, wenn der Beobachtende selbst Teil des Geschehens ist, welches er beobachtet. Für die Unterrichtsbeobachtung ist diese Beobachtungsform von zentraler Bedeutung und wird von Lehrerinnen und Lehrern häufig praktiziert. Die Beliebtheit der teilnehmenden Beobachtungsform unter Lehrpersonen lässt sich nicht zuletzt auch auf die Möglichkeit ihrer flexiblen und kurzfristigen Anwendbarkeit zurückführen. Eine Klasse zu führen und dabei gleichzeitig detailliert und exakt zu beobachten ist jedoch oft nur in eingeschränktem Masse und für kurze Zeit möglich. Ausserdem kann die starke Dominanz der Lehrperson im Unterricht dazu führen, dass sie letztlich etwas beobachtet, das sie zuvor selbst durch ihr Handeln oder ihre Anweisungen provoziert hat (vgl. Köhne 1979, p. 26).

Von einer *nicht-teilnehmenden Beobachtung* kann gesprochen werden, wenn der Beobachter anwesend ist, aber nicht an den Prozessen und Interaktionen einer Gruppe partizipiert, sondern diese von «aussen» beobachtet (vgl. Ingenkamp 1997, p. 58). Im schulischen Umfeld können nicht-teilnehmende Beobachtungen durch Hospitationen ermöglicht werden. Dies ist beispielsweise der Fall, wenn eine Lehrperson den Unterricht einer Kollegin oder eines Kollegen beobachtet. Nicht-teilnehmende Beobachtungen durch Hospitationen entlasten die unterrichtende Lehrperson und fördern, durch die Rückmeldung des Hospitanten, die Reflexion des eigenen Unterrichts und der eigenen Beobachtungsfähigkeit. Nicht-teilnehmende Beobachtungen bergen aber auch die Gefahr reaktiver Effekte in sich. Von einem reaktiven Effekt wird gesprochen, wenn die Beobachtungssituation Einfluss auf das Verhalten der Beobachteten hat (vgl. Lamnek 2001, p. 268).

4.3.2 Offenes vs. verdecktes Beobachten

Offene respektive verdeckte Beobachtungen lassen sich dadurch unterscheiden, ob die Beobachteten wissen, dass sie Objekt einer Beobachtung sind oder nicht. Wissen Personen, dass sie beobachtet werden, handelt es sich um eine offene Beobachtung. Von einer verdeckten Beobachtung kann gesprochen werden, wenn Beobachtete nicht oder zumindest nicht genau wissen, wann und von wem sie beobachtet werden (vgl. Lukesch 1998, p. 123). Diese Beobachtungsform eröffnet dem Beobachter den Vorteil, dass reaktive Effekte mehrheitlich ausbleiben. Verdeckte Beobachtungen können u.a. mit Hilfe eines Einwegspiegels[79], z.B. zur Beobachtung der Interaktion von Kindern, durchgeführt werden. Die Anwendung des verdeckten Beobachtens kann jedoch mit dem ethischen Prinzip des Anrechts auf einen Intimbereich konfligieren.

4.3.3 Standardisierte Beobachtung vs. unstandardisierte Beobachtung

Die Datenerhebung für eine standardisierte respektive gebundene Beobachtung erfolgt mittels eines vorab erstellten Beobachtungskonzepts, welches den Beobachtenden Art und Inhalt der Beobachtung vorgibt. Auch die Anzahl der zu erfolgenden Beobachtungen und die Form der Aufzeichnung und Kategorisierung der Beobachtungsinhalte wird a priori definiert (vgl. Lamnek 2001, p. 267). Im Gegensatz dazu liegt der unstandardisierten Beobachtung kein Schema zugrunde. Als Vorgabe dienen lediglich die leitenden Forschungsfragen. Unstandardisierte Beobachtungen werden häufig mit dem Ziel durchgeführt, Fragestellungen in Bezug auf ein interessierendes Phänomen erstmals zu generieren bzw. sich einem Phänomen in einem ersten Schritt relativ offen zu nähern (vgl. Köhne 1979, p. 25). Die Unterschiede zwischen standardisierten und unstandardisierten Beobachtungen sind jedoch nicht grundsätzlicher, sondern eher gradueller Art.

---

[79] Eine Spiegelwand trennt Beobachtende und zu Beobachtende. Letztere können die Beobachter aufgrund der Verspiegelung des Glases nicht erkennen.

«Man kann sich ein Kontinuum vorstellen, welches von der unstandardisierten bis zur perfekt standardisierten Beobachtung mit einer Reihe von Zwischenformen verläuft. Ein Minimum an Struktur erfordert auch die unstandardisierte Beobachtung und eine gewisse Flexibilität besteht auch bei einer standardisierten Beobachtung» (Lamnek 1995, p. 250).

4.3.4 Kontinuierliche vs. diskontinuierliche Beobachtungen

Eine kontinuierliche Beobachtung ist meistens aus ökonomischen sowie organisatorischen Gründen nicht möglich und auch nicht sinnvoll. Eine Alternative dazu bietet die diskontinuierliche (fraktionierte) Beobachtung, die sich aus Zeitstichproben (time sampling) zusammensetzt. Dabei wird das Verhalten verschiedener Gruppenmitglieder in einer festgelegten Reihenfolge für jeweils beispielsweise drei Minuten beobachtet. Die wesentliche Schwierigkeit einer diskontinuierlichen Beobachtung besteht jedoch darin, gewisse Verhaltensweisen trotz relativ kurzer Beobachtungsdauer erfassen zu können (vgl. Ingenkamp 1997, p. 59).

4.3.5 Labor- vs. Feldbeobachtung

Bei Beobachtungen von Geschehnissen in ihrer alltäglichen Umwelt kann von einer Beobachtung in einer natürlichen Situation oder von Feldbeobachtungen gesprochen werden. Diese verfügen über einen hohen Grad an Natürlichkeit und Alltäglichkeit und sind dadurch komplex und begrenzt kontrollierbar (vgl. Lamnek 2001, p. 268). Im Gegensatz dazu finden Laborbeobachtungen in künstlich konstruierten Situationen statt, die aufzeigen sollen, wie ein Individuum unter dem Einfluss einer bestimmten Variable reagiert (vgl. Köhne 1979, p. 25f.). Beobachtungen in diesen künstlich geschaffenen Situationen ermöglichen eine weitgehende Kontrolle der Verhältnisse aufgrund einer Reduktion der Komplexität. Befunde aus einer künstlich hergestellten Beobachtungssituation sind nach Lamnek «immanent abgesicherter [als Feldbeobachtungen], lassen sich aber nur begrenzt auf natürliche soziale Situationen übertragen» (Lamnek 2001, p. 268).

4.3.6 Medial vermitteltes vs. unvermitteltes Beobachten

Multidimensionalität, simultane Abläufe, Unmittelbarkeit und Unvorhersehbares machen den Unterricht zu einem komplexen Ereignis (vgl. Doyle 1986, p. 394f.). Die Beobachtung desselben ist, trotz einer Eingrenzung des Beobachtungsfeldes, bisweilen anspruchsvoll und setzt ein hohes Mass an Fähigkeiten und Kenntnissen der Beobachtenden voraus. Damit sich die beobachtende Person nicht auf mehrere Aspekte gleichzeitig konzentrieren muss, kann sie zur Beobachtungsunterstützung Aufzeichnungsgeräte (Videokamera, Tonbandgerät) zu Hilfe nehmen (vgl. Lamnek 2001, p. 268). Diese so genannt vermittelte Beobachtung via Medieneinsatz ermöglicht eine dauerhafte Speicherung und eine wiederholbare Abrufbarkeit der Beobachtungsinhalte. Dadurch können beobachtete Sequenzen genauer und wiederholt analysiert werden. Im Gegensatz zur vermittelten Beobachtung werden bei der unvermittelten Beobachtung keine technischen Hilfsmittel eingesetzt. In die Beobachtungsprotokollierung fliessen nur diejenigen Ereignisse ein, die von der Beobachtungsperson bewusst wahrgenommen und verarbeitet werden können. Vergessenes oder nachträglich unklar Erscheinendes lässt sich nach der Beobachtung vielfach nicht mehr schlüssig klären. Im Gegensatz dazu können technische Hilfsmittel diese Lücke schliessen und verhelfen dem Beobachter oder der Beobachterin zu einer Kontrolle der eigenen Beobachtungsqualität (vgl. Abschnitt 4.4.2).

4.3.7 Zusammenfassung

Aus den oben aufgeführten wissenschaftlichen Beobachtungsformen ergeben sich durch verschiedene Kombinationsvarianten, von denen sich einige untereinander logisch ausschliessen, eine Viel-

zahl verschiedener Beobachtungsmöglichkeiten. Die denkbare Vielfalt der beschriebenen Beobachtungsformen wird jedoch letztlich durch die Forschungsrealität begrenzt.

Im schulischen Kontext lassen sich beinahe alle Beobachtungsformen anwenden, wobei einige besser auf den Schulalltag zugeschnitten sind als andere. Die teilnehmende Beobachtung beispielsweise ermöglicht der Lehrperson, ohne personellen Mehraufwand das Unterrichtsgeschehen in der Klasse zu beobachten. Dabei eignen sich sowohl offene als auch verdeckte Beobachtungen. Standardisierte Beobachtungen finden ihre Anwendung in erster Linie dann, wenn es darum geht, Informationen zu erhalten, die später in eine schriftlich abgefasste Beurteilung einfliessen sollen. Wie im Fallbeispiel gezeigt, entdecken Lehrpersonen aber vielfach zufällig Dinge, die das Interesse an einer zu Beginn meist unstandardisierten Beobachtung wecken. Laborbeobachtungen finden im schulischen Kontext kaum und technisch vermittelte Beobachtungen selten Anwendung.

Auch wenn Lehrerinnen und Lehrer in ihrer Praxis in der Regel nicht auf alle der vorgängig erwähnten Formen der Unterrichtsbeobachtung zurückgreifen können und eine ausgeprägte Systematisierung der Beobachtungen, wie sie in der Wissenschaft verfolgt wird, selten möglich ist, sollte es trotzdem Ziel der Lehrpersonen sein, Beobachtungen innerhalb der zur Verfügung stehenden Ressourcen systematisch durchzuführen.

4.4 Formen der Beobachtungsprotokollierung

Aufgabe der Protokollierung ist das Festhalten eines flüchtigen Moments, um ihn mittelbar, d.h. zum Gegenstand, zu machen. Das Protokollieren bildet ein Grundgerüst von qualitativen oder quantitativen Daten zur zielgerichteten und methodischen Auswertung. Damit erhalten Protokollierungstechniken ein zentrales Gewicht in einem Beobachtungsinstrument.

4.4.1 Protokolltechniken

Protokolltechniken unterscheiden sich in der Art des Festhaltens und der Strenge der Strukturierung. Beobachtungen können mit Strichen, Zahlen, Zeichen, Graphiken, Stichwörtern oder ausgeschriebenen Sätzen schriftlich festgehalten werden. Der Einsatz von audiovisuellen Hilfsmitteln bildet eine mögliche Entlastung bei der Erhebung der Daten.

4.4.1.1 Gedächtnisprotokoll

Kleber (1992, p. 201) nennt die Gedächtnisprotokolle die am häufigsten verwendete Form des Festhaltens von Beobachtungen. Bei dieser Protokollierungsform wird nach relevanten Ereignissen das Geschehene schriftlich notiert. Die Beobachtungen können in Sätzen oder in Stichworten dokumentiert werden und erfolgen meistens ohne schriftlich fixierte Beobachtungsrichtlinien. Gedächtnisprotokolle finden ihre Anwendung oft nach längeren zeitlichen Abschnitten und grösseren Unterrichtseinheiten, wie zum Beispiel nach einem Schultag oder am Ende eines Seminars. Ähnliche Formen der Protokollierung sind das narrative Protokoll, das naive Protokoll, das Verlaufsprotokoll, die Zufallsbeobachtung, das Erlebnisprotokoll, die anekdotische Beobachtung oder die freie Protokollierung. Bekannt ist das pädagogische Tagebuch, bei dem relevante Beobachtungen vom Tage in einer Nachbesinnung strukturiert und regelmässig aufgezeichnet werden.

Dieses Vorgehen ist aufgrund des zufälligen Charakters der Beobachtungen und hohen Grads an Subjektivität problematisch. Vorteile können sich in der Initiation von neuen Fragestellungen und gezielteren Beobachtungen ergeben. Mitunter stellen sich interessante Fragen erst im Verlauf eines Gedächtnisprotokolls. Das Vorgehen entspricht dann dem Stil einer explorativen Studie: Forschungsziel ist nicht die Prüfung einer Hypothese, sondern deren Entwicklung anhand von beobach-

tetem und gesammeltem Material. Sinnvoll werden Gedächtnisprotokolle damit vor allem zu Beginn einer Diagnosephase und sollten später durch systematische Verfahren ergänzt werden.

Verzerrungen beim Gedächtnisprotokoll können sich aufgrund erinnerungsbezogener Veränderungen der Informationen ergeben. Einige Beobachtungen werden zu komplexen Einheiten zusammengefasst, andere vergessen oder verdrängt, vereinfacht oder reduziert. Vor allem die negativen Auffälligkeiten und die Interpretationen bleiben bei der Niederschrift im Gedächtnis. In der Erinnerung werden Beobachtungen an implizite Theorien angepasst. Kleber (1992, p. 202) schlägt als Gegenmassnahme einen «Erinnerungszwang zum Konkreten» vor: Interpretationen sollten zur Protokollierung in konkret Beobachtetes zurück übersetzt werden. Die beobachtende Person sollte gesehenes Verhalten bewusst verbalisieren, damit die Erinnerungen konkreter behalten werden. Und Gelegenheitsbeobachtungen sollten so schnell wie möglich aufgeschrieben werden, um die Verzerrungen durch spätere Deutungen zu reduzieren.

4.4.1.2 Strichlisten

Strichlisten bieten sich an, wenn Hypothesen des zu Beobachtenden so weit spezifiziert sind, dass die Häufigkeiten von Verhaltensweisen erfasst werden können. Der interessierende Aspekt fokussiert sich auf einige wenige Verhaltensweisen. Damit ergibt sich eine höhere quantitative Präzision, andere relevante Bedingungen und Merkmale hingegen werden nicht registriert. Wichtig bei Strichlisten ist, dass die in der Planung bestimmten und strukturierten Verhaltensweisen eindeutig zu den Kategorien zugeordnet werden können.

Sind die interessierenden Merkmale einfach zu beobachten, kann die Lehrperson im Rahmen einer teilnehmenden Beobachtung selber die Registrierung vornehmen. Verdeckte oder häufig auftretende Merkmale erfordern hingegen den Einbezug eines Fremdbeobachters oder einer Fremdbeobachterin (vgl. Wahl et al. 1997, p. 295f.). Verdeckt ist ein Verhalten, wenn Beobachtende aus offenen Merkmalszusammenhängen und Symptomen (zum Beispiel «lachen» oder «gähnen») die interessierenden Variablen (zum Beispiel «freut sich» oder «ist müde») ableiten und dadurch bereits eine Interpretationsleistung erbringen müssen. Fremdbeobachtung ermöglicht bei einer kleinen Zahl von Zeilen und Spalten in einem übersichtlichen Beobachtungsbogen ein Protokollieren über einen längeren Zeitraum hinweg. Bei teilnehmender Beobachtung sind längere Intervalle der Protokollierung mit der Methode der Strichliste nicht praktikabel und schwer durchführbar.

Beispiel einer systematischen Erhebung mittels Strichlisten ist das Interaktiogramm (vgl. Abbildung 4.1). Im Rahmen einer Gruppendiskussion wird in einer Matrix mit Strichen erfasst, wer wen anspricht. Dynamiken einer Gesprächsgruppe können sich dann im Zusammenrechnen der markierten Striche ergeben. Die vertikale Summe der Matrix zeigt die Häufigkeit der verbalen Zuwendung eines Mitglieds an andere Mitglieder oder an die Gesamtgruppe, die horizontale Summe die Häufigkeit des «Angesprochenwerdens» (vgl. Köhne 1979, p. 35f.). Resultate von Strichlisten sind quantitativer Natur, die für Fragestellungen mit dem Anspruch auf qualitative Beobachtungsbefunde oft keine interpretierbaren Daten liefern. Ein Kriterium für die Anwendung von Strichlisten ist die Aussicht eines schnellen Erkenntnisgewinns für die Lehrperson.

Diskussionsteilnehmer pass. / akt.	A	B	C	D	E	F	G	Gruppe	Σakt.	Rang akt.
A	—									
B		—								
C			—							
D				—						
E					—					
F						—				
G							—			
Σpass.										
Rang pass.										—

Abbildung 4.1: Schema eines Interaktiogramm-Bogens zur Registrierung verbaler Beziehungen in einer Gruppendiskussion (aus: Köhne 1979, p. 36)

4.4.1.3 Wortprotokoll

Beim Wortprotokoll werden Äusserungen, Merkmale und Verhaltensweisen von Schülerinnen und Schülern sowie von Lehrpersonen wörtlich erfasst. Diese können in Form von Schilderungen, Beschreibungen, Darstellungen oder Vergleichen festgehalten werden. Wortprotokolle werden meist mit einer Zeitleiste ergänzt, zum Beispiel mit einer Kennzeichnung der Zeitabschnitte alle 10 Minuten. Die einzelnen Redebeiträge können für eine übersichtlichere Auswertung chronologisch nummeriert werden. Film- und Tonbandaufnahmen, sofern sie nicht technische Mängel aufweisen, können eine hilfreiche Ergänzung sein.

Im Unterricht entstandene Wortprotokolle sind infolge ihrer aufwändigen Protokollierung in der Regel fraktionierte Beobachtungen (time sampling), d.h. zeitlich begrenzt. Für Erinnerungen eignen sich zum Beispiel Stichwortprotokolle, wobei bei der Protokolltechnik nur der Schreiber oder die Schreiberin selbst das Notierte sinngemäss assoziieren und folglich verstehen kann. Werden Wortprotokolle während des Unterrichts in ausgeschriebenen Sätzen dokumentiert, braucht es wegen des Aufwands in der Anwendung einen geübten Fremdbeobachter, unter Umständen mit stenographischen Fähigkeiten.

Wortprotokolle sind bei den Erhebungen meistens niedrigsystematisierte und niedriginferente[80] Verfahren, welche im Rahmen der Auswertung kodiert werden. Als Aufnahmeverfahren bleiben sie offen für Hypothesenbildung und erlauben einen besseren Blickwinkel, um kontextuelle Einflüsse von Verhalten zu erfassen. Da es nicht möglich ist, jedes Detail festzuhalten, sind die Beobachter auch hier dazu aufgefordert, die Aufnahme von Informationen auf die relevanten Fragen der Untersuchung zu verringern. Wortprotokolle benötigen ebenso ein Minimum an Regeln. Deutungen, Wertungen und Urteile der Protokollierenden sollten nicht protokolliert werden, « […] da durch sie als gültig unterstellt würde, was mit Hilfe der Beobachtung erst geprüft werden soll» (Mollenhauer & Rittelmeyer 1977, p. 150).

[80] Das Inferenzniveau einer Beobachtung (hochinferent, niedriginferent) hängt davon ab, wie viele Informationsverarbeitungsschritte der Beobachtende leisten muss.

4.4.1.4 Teilformalisiertes Protokoll

Ein teilformalisiertes Protokoll siedelt sich zwischen alltäglicher und wissenschaftlicher Beobachtung an. Alltagsbeobachtung ist situativ, zufällig, unsystematisch und von persönlichen Interessen geprägt. Wissenschaftliche Beobachtung ist hypothesengeleitet, systematisch, strukturiert, standardisiert, dokumentiert und ausgewertet. Das teilformalisierte Protokoll kann mit einer begutachtenden Beobachtung verglichen werden: Sie ist zielbestimmt, kriterienorientiert und richtet ihren Fokus auf einen bestimmten Ausschnitt oder Aspekt. Teilformalisierte Protokolle sind für Unbeteiligte unter Umständen nur bedingt nachvollziehbar.

Die unterschiedlichen Protokolltechniken sind in den Anwendungen nicht an ihre reine Form gebunden. Systematische Unterrichtsbeobachtungen können mehr oder weniger formalisiert werden, mit den entsprechenden Vor- und Nachteilen. Lukesch (1998, p. 117) nennt die Frage nach der Eingrenzung von Verhaltensmerkmalen in Kategoriensystemen das «Breitband-Fidelitäts-Dilemma». Je konkreter Beobachtungsinstrumente bestimmte Verhaltensbereiche oder Persönlichkeitsmerkmale erfassen, desto kleiner wird die Validität der Aussage. Relevante Bereiche der Diagnostik können unbeachtet bleiben oder verloren gehen. Zudem lassen sich im Geschehen des Unterrichts die interessierenden Verhaltensweisen und Merkmale nicht immer beobachten. Ein weiterer Aspekt, den es zu beachten gilt, ist die Frage der Machbarkeit. Viele Beobachtungsverfahren sind in wissenschaftlichen Kontexten entstanden und für den Gebrauch im Unterricht zu aufwändig. Ein komplexes Erhebungsinstrument erschwert einer Lehrperson gleichzeitiges Unterrichten und Beobachten.

4.4.2 Technische Hilfsmittel

Technische Hilfsmittel werden eingesetzt, um medial vermittelte Beobachtungen durchzuführen. Der Einsatz von audiovisuellen Hilfsmitteln kann den Beobachter bei der Erhebung von Daten entlasten. Die technischen Hilfsmittel sind im Wesentlichen Video- und Tonaufnahmen, welche nach den Aufzeichnungen ausgewertet werden. Die Videoaufzeichnung gilt mitunter als die zuverlässigste Methode der Protokollierung von Beobachtungen. Ein auf Video «konservierter» Unterricht kann durch Wiederholungen, Verstärkertechnik, Zoomaufnahmen und eingeblendeten Ausschnitten helfen, die Beobachtungsleistungen zu verbessern. Ein pädagogischer Nutzen liegt in der methodischen Möglichkeit, die Aufzeichnungen vorzuführen. Gemeinsam mit den Beobachteten können Analysen und Diskussionen durchgeführt werden. Der Beobachtete wird damit zum Beobachter seiner selbst. Technische Hilfsmittel bieten sich zum Beispiel bei grösseren Gruppen an, wo eine unvermittelte Unterrichtsbeobachtung nicht problemlos möglich ist.

Einschränkungen von Video- und Tonbandaufnahmen liegen in ihrem begrenzten Fokus. Beim Film ergibt sich der Bildausschnitt aus dem Standpunkt der Kamera, die im Gegensatz zum Beobachter kein «Weitwinkelobjektiv» besitzt. Wird die Kamera nicht von einem Kameramann bedient, bleibt sie an einem Ort fixiert. Gleich verhält es sich bei den Tonbandaufnahmen. Die Position des Mikrophons bestimmt den akustischen Ausschnitt. Ein weiterer spezifischer Nachteil kann in der technischen Qualität der Aufnahmen liegen. Weiter weg liegende Geräusche können bei Tonaufnahmen von nahen Geräuschen überlagert werden, was zu einem Verwischungseffekt des Klangbildes führt. Störeinflüsse können sich auch durch Schalleffekte des Raums ergeben. Bei Tonaufnahmen ist die Tonaussteuerung eine Aufgabe, die Wissen und Übung voraussetzt und auch nur mit den dazu nötigen Geräten möglich ist. Probleme bei der Beobachtung von eindeutigem Verhalten aufgrund von Filmmaterial können sich zum Beispiel bei schwacher Beleuchtung oder Unschärfe der Aufnahmen ergeben.

Der Einsatz von Medien provoziert verändertes Verhalten im sozialen Feld (vgl. Lamnek 2001, p. 269). Vor allem am Anfang des Einsatzes von Aufzeichnungsgeräten kann dies störend auf die Klassensituation wirken, wobei sich in der Regel ein rascher Gewöhnungseffekt einstellt. Für Schüler und Schülerinnen sollten die Aufzeichnungen wissentlich und transparent stattfinden. Aufzeich-

nungsgeräte dürfen nicht als «Überwachungsinstrumente» eingesetzt werden. Bester Schutz gegen Artefakte von Aufnahmen ist die informierte Mitwirkung aller Beteiligten. Aufzeichnungen von Beobachtungen mit Hilfe von technischen Hilfsmitteln, auch mit Aufklärung der Betroffenen, sind ethische Grenzen gesetzt. Sie tangieren den Datenschutz verstärkt, weil die Anonymität der Aufzeichnungen infolge identifizierender Bilder und Töne schwerer gewährleistet werden kann. Die Daten sind vertraulich und dürfen nicht ohne Zustimmung an Dritte weitergegeben werden (vgl. Ziefuss 1978, p. 101f.).

Die Auswertung von audiovisuellen Aufzeichnungen erfordert viel Zeit. Um die Aufzeichnungen ergiebig auswerten zu können, empfiehlt sich auch hier als einer der ersten Schritte die Aufstellung einer systematischen Beobachtungsmethodik mit definierten Beobachtungskategorien. Bei der Transkription (Verschriftlichung der Bandaufzeichnung) und der Dokumentation von Filmaufzeichnungen erleichtert die Auswahl kurzer Sequenzen die Analyse.

4.5 Beobachtungssysteme

Beobachtungssysteme werden in der Literatur hauptsächlich in drei Arten unterschieden: Zeichensysteme, Kategoriensysteme und Ratingskalen. Auch Soziogramme sollen – wenn auch in einer leicht modifizierten Form – in diesem Kapitel kurz besprochen werden. Die Differenzierungen bilden, wie auch die Unterscheidungen der Protokollierungstechniken, keine sich ausschliessenden Klassen. In einem Beobachtungsschema können verschiedene Systeme zur Anwendung kommen (vgl. Ziefuss 1978, p. 125ff.; Kleber 1992, p. 203; Ingenkamp 1997, p. 60ff.; Wahl 1997, p. 288; Lukesch 1998, p. 129f.; Lamnek 2001, p. 273f.).

4.5.1 Überblick über die Beobachtungssysteme

In diesem Abschnitt werden zuerst Zeichensysteme, dann Kategoriensysteme und am Schluss Ratingskalen dargestellt und illustriert.

4.5.1.1 Zeichensysteme

Zeichensysteme oder Merkmalsysteme basieren auf der Beobachtung von genau definierten Verhaltensmerkmalen, darstellt als Items. Die Auswahl der Indikatoren zur Konstruktion eines strukturierten Beobachtungsprotokolls erfolgt nach denkbaren Verhaltensmerkmalen, welche für einen bestimmten Bereich sozialen Geschehens als symptomatisch angesehen wird. Diese Beschränkung hat zur Folge, dass ein grosser Teil der Beobachtungen nicht festgehalten wird.

Die in der Vorbereitung interessierenden Merkmale werden nach der Häufigkeit ihres beobachteten Auftretens in einem Beobachtungszeitraum registriert. Die notierten Verhaltensäusserungen erfordern in der Regel keine Bearbeitung, d.h. keine Kodierung. Der zeitliche Rahmen einer Beobachtungsphase ist meistens relativ kurz, zwischen 3 und 5 Minuten. Erfasst werden Verhaltensmerkmale wie zum Beispiel «Schüler schlägt Mitschüler». Dabei können mehrere Merkmale gleichzeitig auftreten, wie «Schülerin schlägt Mitschüler» und «Schülerin flucht». Weil es beim Zeichensystem wenig zu protokollieren gibt, erlauben sie die Beobachtungen vieler Verhaltensvarianten. Dies setzt voraus, dass die einzelnen Merkmale eindeutig und verständlich den Kategorien zugeordnet werden können.

Nicht geeignete Formulierungen zur Kategorisierung im Zeichensystem sind abstrakte, vom Verhalten abgehobene Beschreibungen, wie zum Beispiel «autoritärer Erziehungsstil». Eine solche Kategorie wäre für die Durchführung einer Beobachtung in der Regel nicht brauchbar. Praktikable Kategorien bilden konkretere Verhaltensweisen wie «Sprechzeiten von Lehrpersonen sowie Schüle-

rinnen und Schülern» oder «Einsatz von Verstärkern», d.h. Lob oder Tadel, protokolliert nach Häufigkeit ihres Auftretens.

Das bekannteste und in der Literatur am meisten aufgeführte Zeichensystem ist OScAR (Observation Schedule and Record, Schema zur Beobachtung und Aufzeichnung) von Medley et al. (1966) (vgl. Ziefuss 1978, p. 191ff.). Praktisches Beispiel eines Zeichensystems ist der «Beobachtungsbogen zur Kommunikation in Kleingruppen in der Grundschule» (Herrmann & Höfer 1999, p. 76). Thema ist das Einhalten von Gesprächsregeln, welche in diesem Beispiel (vgl. Abbildung 4.2) die Spalten definieren. Während der Beobachtung wird ein Strich in der Spalte notiert, wenn eine Schülerin oder ein Schüler das betreffende Verhalten zeigt. Die Auswertungen können summativ, d.h. abschliessend durchgeführt werden.

Beobachtungsbogen (für den Lehrer)

3. Schj. SU Kleingruppe (4 K.)

Schwerpunkt: Fragen zu einem bekannten Text stellen

Name	Blick-kontakt	Hand-zeichen	deutlich	zuhören	ausreden	Variationen

Abbildung 4.2: Beobachtungsbogen zur Kommunikation in Kleingruppen (aus: Herrmann & Höfer 1999, p. 77)

4.5.1.2 Kategoriensysteme

Kategoriensysteme unterscheiden sich von den Zeichensystemen insofern, als sie nicht einzelne Verhaltensmerkmale, sondern Verhaltenseinheiten zu erfassen versuchen. Jede Verhaltenseinheit wird in eine vorgeschriebene Kategorie klassifiziert. Das erfordert eine andauernde Bearbeitung und Protokollierung des Beobachteten, d.h. eine Entscheidung über die Zuordnung im Kategoriensystem. Beobachtet werden können alle Verhaltensweisen in einem begrenzten Zeitraum, womit man formal ein vollständiges Protokoll des Geschehens erhält, oder fokussiert auf bestimmte Verhaltensbereiche, die in Zusammenhang mit der Fragestellung relevant sind. Kategoriensysteme beziehen sich meistens auf verbale Äusserungen (Ziefuss 1978, p. 129).

Der Entwicklung eines Kategoriensystems kommt eine zentrale Bedeutung bei der Durchführung einer Beobachtung zu. Die Kategorien repräsentieren die interessierenden Variablen der zentralen Fragestellung. Die Güte des Beobachtungsinstruments wird entscheidend von der Genauigkeit und der Sorgfalt, mit der die Kategorien beschrieben und ausgewählt werden, bestimmt. Beobachtungskriterien sollten nicht aufgrund von Zufälligkeiten, Willkürlichkeiten, persönlichen Präferenzen oder Erwartungshaltungen erstellt werden, sondern auf Basis von Theorien, vorhandenen Forschungsergebnissen und persönlichen Einsichten. Die beobachtbaren Indikatoren (Merkmale) werden in der Planung bestimmt und kategorisiert.

Formale und inhaltliche Kriterien zur Konstruktion eines Kategoriensystems sind:

- Eindimensionalität (ein Merkmal sollte nicht mehreren Teilmerkmalen entsprechen);

- Ausschliesslichkeit (ein Ereignis darf nur einer Kategorie zugeordnet werden können);
- Vollständigkeit (alle relevanten Beobachtungen müssen erfasst werden können);
- Eindeutigkeit (die beobachtbaren Sachverhalte müssen den Kategorien zugeordnet werden können);
- Begrenzung (nicht mehr Kategorien, als der Beobachter wahrnehmen und behalten kann).

Die Anzahl der Kategorien sollte auf eine beobachtbare und interpretierbare Dimension beschränkt sein. Infolge der kleinen Kategorienzahl wird jede Kategorie normalerweise mehrmals zur Protokollierung der verschiedenen beobachteten Verhaltensmerkmale herangezogen. Im Gegensatz zum Zeichensystem ist beim Kategoriensystem ein gleichzeitiges Zuordnen und Protokollieren von einem Verhaltensbereich zu zwei Kategorien nicht möglich.

Kategoriensysteme liefern mehr Daten und sind flexibler als Zeichensysteme. Erfolgen unterschiedliche Erhebungen nach einheitlichen Kategorien, erlauben sie einen Austausch und Vergleichbarkeit. Infolge des höheren Abstraktionsniveaus und der grösseren Anzahl von schlussfolgernden Schritten beim Beobachter sind aber die Resultate weniger verlässlich (vgl. Ziefuss 1978, p. 129). Wie viel Inferenz während des Beobachtungsprozesses nötig ist, hängt von den gewählten Kategorien und Merkmalen ab. Bei niedriginferenten Beobachtungen werden zum Beispiel die Häufigkeiten des Auftretens von Schimpfwörtern bei Schülerinnen und Schülern gezählt, hochinferent sind Aussagen wie «Schüler ist aggressiv». Hochinferente Urteile können zu unterschiedlichen Wertungen führen, weshalb die Beurteilenden eine genaue Beschreibung der Bedeutung der Merkmale und der Kategorien angeben sollten. Konkrete Beispiele zu den Kategorien erhöhen zusätzlich die Transparenz. Das gleiche Schülerverhalten interpretiert und wertet ein Beobachter als «Egoismus», für einen anderen Beobachter ist es «Durchsetzungsvermögen». Als Beispiel für verhaltenskategorisierende Instrumente sei eines der bekanntesten Kategoriensysteme genannt: Das System von Robert Bales zur Beschreibung von Interaktionen kleiner Gruppen (vgl. Fassnacht 1995, p. 184). Dazu Bales (1967, p. 152, zit. n. ebd.):

> «Jede Kategorie ist sowohl im Hinblick auf alle anderen Kategorien [...] definiert. Die Reihe dieser Kategorien gilt als umfassender und systematischer als eine blosse Liste. Man behandelt sie so, als würde sie auf ihrer eigenen Abstraktionsebene logisch alle Möglichkeiten erschöpfen. Man nimmt an, dass jede mögliche Handlung irgendeiner der 12 Kategorien richtig zugeordnet werden kann. Ferner sind alle Kategorien positiv definiert, das heisst, keine einzige wird als Restkategorie betrachtet, unter der man alles subsumiert, was sonst nicht untergebracht werden kann. Jede Handlung wird nur einer einzigen positiven Kategorie zugeordnet.»

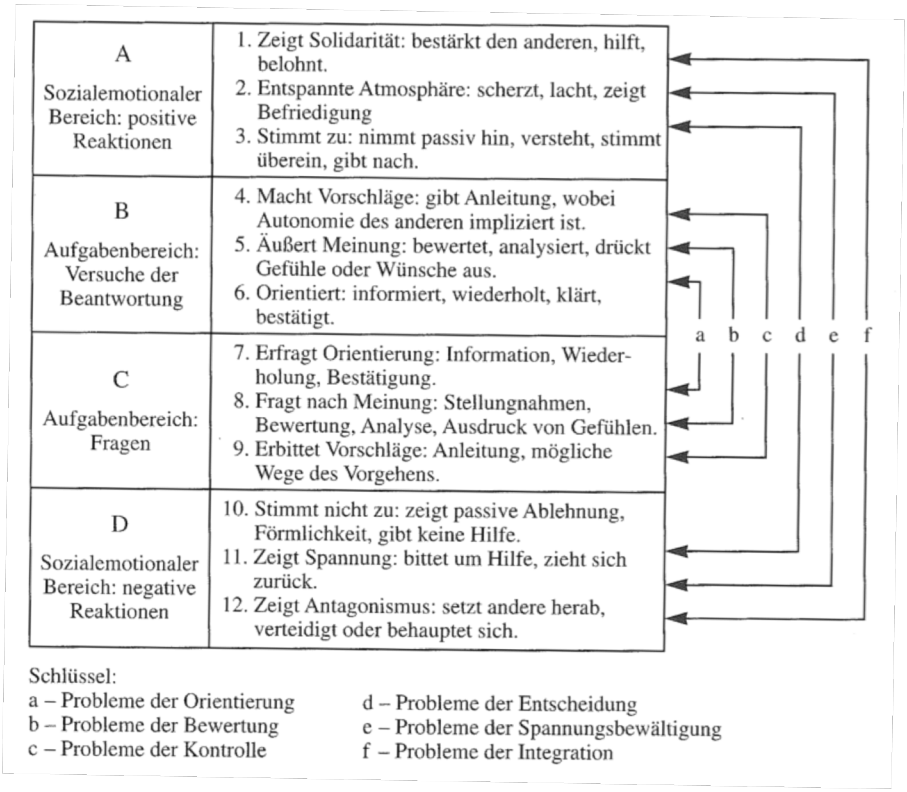

Abbildung 4.3: Kategorien-System «Interaktion kleiner Gruppen» nach Robert Bales (1967, zit. n. Fassnacht 1995, p. 184)

4.5.1.3 Ratingskalen

Ratingskalen oder Schätzskalen beruhen auf Einschätzungen des Beobachtenden. Nach bestimmten Merkmalen wird das beobachtete Verhalten quantifiziert, d.h. die Beobachtung wird in Zusammenhang mit dem Grad der Ausprägung von Merkmalen gestellt. Die Beobachterin oder der Beobachter legt sowohl das Messinstrument wie auch den Beurteilungsmassstab selber fest. Beurteilungen erfolgen im Sinne eines gradmässigen Vergleichens und dem Erstellen einer Rangordnung. Die Rangordnung bedarf dabei keiner exakten Messung, eindeutigen Regeln oder Indikatoren als Prüfitems. Was unter dem eingeschätzten Urteil dann genau zu verstehen ist, bleibt meistens den Beobachtenden überlassen. Die Beurteilungen können nach einem oder mehreren Gesichtspunkten erfolgen, das Resultat bildet ein globales Urteil. Dabei handelt es sich meistens um Beurteilungsskalen aufgrund von Eigenschaften auf hochinferenter Ebene. Ratingurteile sind oft «mehrdimensional», d.h. verschiedene Beobachtende kommen bei der gleichen Beobachtung zu verschiedenen Ergebnissen.

Ingenkamp (1997, p. 65) schlägt gegen solche «spontanen» und «intuitiven» Ratingskalen eine Revision im Sinne möglichst präziser Konzeptionen und konkreter Orientierung vor. Eine abstrakte Beobachtungskategorie wie «Beziehungsfähigkeit des Schülers» mit den Merkmalsausprägungen «selten-gelegentlich-oft-keine Aussage» sind mit Indikatoren zu ergänzen, welche die Beobachtungen konkretisieren. Eine mögliche nähere Bestimmung von der «Beziehungsfähigkeit des Schülers» könnte heissen: «Nimmt von sich aus mit seinen Mitschülern Kontakt auf» (Nuding 1997, p. 21). Im Unterschied zu Zeichen- und Kategoriensystemen ergeben sich bei den Ratingskalen grössere zeitliche Verschiebungen der Protokollierung, welche zu Verzerrungen infolge von Erinnerungsleistungen führen können.

Die Zuordnung des Ausprägungsgrades eines beobachteten Verhaltens kann mit einer Zahl, Graphik oder einer verbalen Bestimmung, z.B. «schwach-mittel-stark», geschehen. Zur Einstufung soll-

ten nicht mehr als 2-5 Unterscheidungen vorgenommen werden. Mehr numerische oder grafische Stufen führen normalerweise nicht zu einer exakteren Messung, fordern aber die Beobachter ertragslos mehr. Nuding (1997, p. 18ff.) unterscheidet im schulischen Kontext zwischen «curriculumbezogenen Ratings» zur Leistungsmessung und «persönlichkeitsbezogenen Ratings» zur Beurteilung von Verhaltenseigenschaften. Üblich ist die Kombination der Ratingskalen mit Zeichen- und Kategoriensystemen. Wie beim Kategoriensystem stellt das Rating-Verfahren während der Observation aufgrund der notwendigen Bewertungs- und Interpretationsleistungen hohe Anforderungen an die Beobachtenden.

Abbildung 4.4 zeigt ein Beispiel einer quantitativ zu schätzenden Verhaltenskategorie aus einem Beurteilungsbogen zur Feststellung der Lernausgangslage von Schulanfängerinnen und -anfängern (vgl. Köhne 1979, p. 38f.).

1. Persönlichkeitsbereich				
1.1. Kontaktfähigkeit: Ist das Kind aktiv im Kontakt? Knüpft es schnell und ungezwungen Beziehungen? Oder ist es eher passiv, scheu, gehemmt im Umgang mit anderen?				
Das Kind zeigt Kontaktfähigkeit:				
sehr häufig	häufig	gelegentlich	äusserst selten	keine Angaben möglich
Erläuterungen:				

Abbildung 4.4: Beispiel einer quantitativ zu schätzenden Verhaltenskategorie (aus: Köhne 1979, p. 38)

Operationalisierungsbeispiele zur besseren Illustration könnten bei der Kategorie «Kontaktfähigkeit» (vgl. Abbildung 4.4) die folgenden sein: «Man sieht das Kind häufig mit anderen reden; es macht ihm offensichtlich nichts aus, andere anzusprechen; es holt sich Material vom Hausmeister; es fragt den Pädagogen um Rat. Oder: Das Kind spielt meisten für sich allein; es wirkt verlegen, wenn man es anspricht; es wirkt unsicher, wenn man ihm eine mit Kontaktaufnahme verbundene Aufgabe gibt (z.B. in einer anderen Klasse etwas holen)» (Köhne 1979, p. 38).

4.5.1.4 Soziogramme

Soziometrische Daten werden neben der Protokollierung aufgrund von Beobachtungsdaten meistens mit der Methode des Wahlverfahrens erhoben. Dabei werden die Schülerinnen und Schüler gebeten, positive Wahlen und allfällige Ablehnungen nach einem bestimmten Kriterium auszuwählen und diese spontan auf einen «Wahlzettel» zu schreiben. Petillon (1980) nennt vier Strukturen mit ihren entsprechenden Kriterien:

- Das Wahlkriterium des Sitznachbarn oder der Sitznachbarin gibt Aufschluss über die *Beliebtheitsstruktur* innerhalb einer Schulklasse.
- Will man Hinweise auf die *Entscheidungsstruktur* einer Klasse erhalten, eignet sich als Wahlkriterium die Wahl einer Klassenvertretung, welche die Interessen der Klasse vertreten soll.
- Schüler in projektbezogene Funktionen wählen lässt (z.B. Rollen in einem Theaterstück). Schliesslich lässt sich durch die Wahl einer Diskussionsleitung die *Kommunikationsstruktur* einer Klasse bestimmen.

- Jedes Wahlkriterium wird anhand einer spezifischen relationalen Frage (Wer-Wen-Frage) erhoben. Jede Schülerin, jeder Schüler muss wählbar sein und selber wählen können. Durch Soziogramme können sowohl dyadische (zwischen zwei Personen) als auch polyadische Beziehungen (zwischen mehreren Personen) in Freundschafts- oder Feindschafts-beziehungen oder Subgruppen erhoben werden[81].

Die sozialen Konsequenzen innerhalb einer Klasse aufgrund einer solchen Befragung (v.a. aufgrund von negativen Wahlen resp. Ablehnungen) dürfen nicht unterschätzt werden. Dies spricht dafür, Befragungen für Soziogramme ausschliesslich von Experten (z.B. Mobbingexperten) durchführen zu lassen. Es gibt aber auch die Möglichkeit, anstelle einer Schülerbefragung Beobachtungen durchzuführen und als Lehrperson anhand dieser Eindrücke selber ein Soziogramm zu machen. Hierbei können nicht alle von Petillon (1980) vorgegebenen Schritte durchgeführt werden, es handelt sich deshalb nicht um ein Soziogramm im strengen Sinne. Aber es resultiert ein Soziogramm-ähnliches «Gebilde» (vgl. Abb. 4.5), das der Visualisierung der sozialen Interaktionen innerhalb einer Schulklasse oder einer Gruppe dient.

Bevor eine Lehrperson die Beobachtung des sozialen Gefüges einer Schulklasse startet, sollte sie sich für bestimmte im Voraus festgelegte Beobachtungskriterien entscheiden. Sie überlegt sich zum Beispiel Fragen, die Antworten auf Beziehungsprobleme geben können und benutzt diese als Beobachtungskriterien.

Man kann sich für die Beobachtung z.B. folgende Fragen stellen:

- Wer spielt mit wem (z.B. in der Pause, vor und nach der Schule oder auf dem Schulausflug)?
- Wer belästigt wen (z.B. in der Pause, vor und nach der Schule oder auf dem Schulausflug)?
- Wer ignoriert wen (z.B. in der Pause, vor und nach der Schule oder auf dem Schulausflug)?
- Wer hilft wem bei den Aufgaben?
- Wer möchte mit wem eine Gruppenarbeit machen?
- Wer möchte mit wem auf keinen Fall eine gemeinsame Aufgabe erledigen?
- Wer möchte neben wem sitzen?

Erst solche Kriterien (Fragestellungen) erlauben die Messung der Beziehungen und Interaktionen *zwischen* den einzelnen Mitgliedern und geben Aufschluss über das Beziehungsgefüge und die Sympathiestruktur innerhalb der Gruppe. Valkanover et al. (2004) schlagen zur besseren Visualisierung von systematischen Beobachtungssequenzen eine grafische Darstellung vor (vgl. ebd., p. 55). In der Annahme, dass Mobbing nicht aufgrund einzelner Verhaltensweisen bestimmt werden kann, soll die systematische Beobachtung über einen Zeitraum von mindestens zwei Wochen dauern. Der Fokus der Lehrperson hat sich auf vier Beobachtungskriterien zu richten: subtile Formen von Gemeinheiten, systematische Handlungen von Mobbing, Ungleichheiten zwischen Opfer und Täter und den Reaktionen aller Beteiligten. Täglich soll die Lehrperson in einer von ihr gewählten Zeit, zum Beispiel während einer bestimmten Pause, so viele Informationen und Einzelheiten wie möglich protokollieren. Zur einfacheren Aufzeichnung dient eine Beobachtungsskizze.

---

[81] Die Aufbereitung solcher Daten anhand einer Soziomatrix sowie deren Auswertung und Visualisierung mit einem Soziogramm werden in Petillon (1980) detailliert beschrieben.

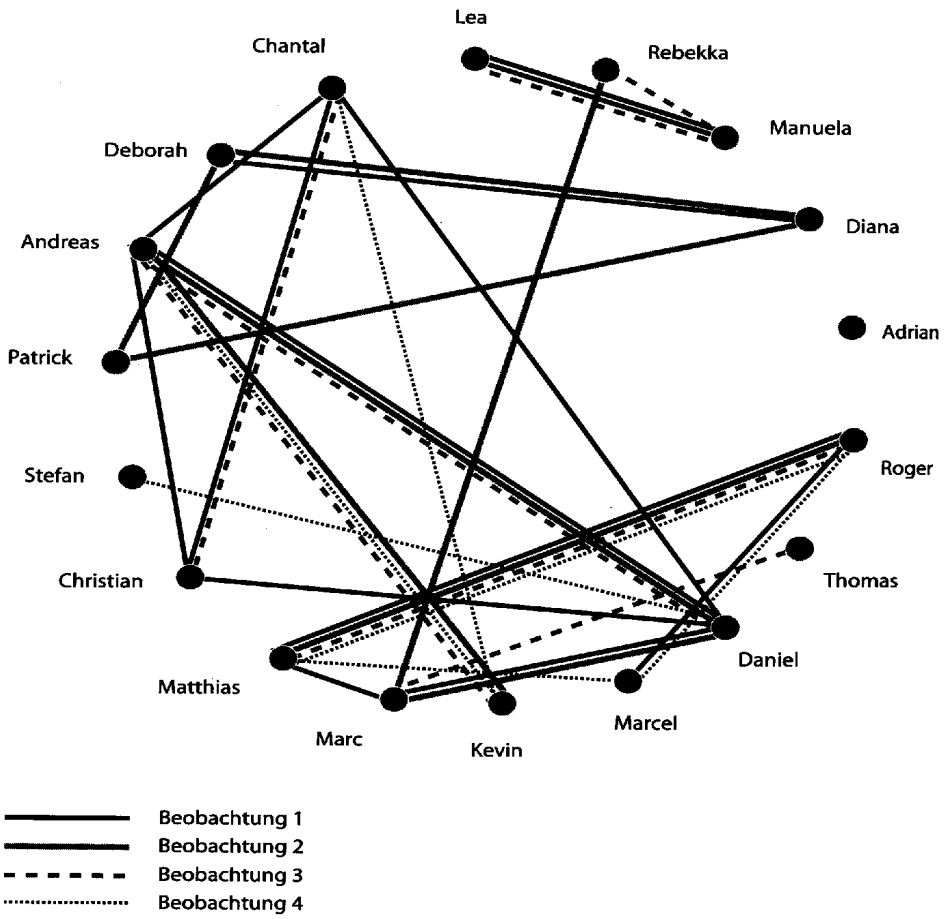

Abbildung 4.5: Beispiel einer Beobachtungsskizze (aus: Alsaker 2004, p. 55)

Zwischen den Personen werden verschieden gekennzeichnete Linien, welche exemplarisch für bestimmtes Verhalten stehen, gezogen. Eine durchgezogene Linie zwischen Kind A und Kind B der kreisförmig aufgestellten Teilnehmer kann heissen: «A belästigt B». Ergebnis ist eine grafische Übersicht, welche zum Beispiel isolierte Kinder ohne Initiierungen und Reaktionen aufdeckt (vgl. Abb. 4.5: Adrian). Besondere Beobachtungen werden gesondert und schriftlich notiert.

Durch solche Visualisierungen erhält die Lehrperson die Möglichkeit, Spannungen, Schwierigkeiten und Vorurteile zu erkennen, um anschliessend lenkend in den Prozess einzugreifen und mit gezielt eingesetzten pädagogischen Massnahmen die Situation für einzelne Gruppenmitglieder oder die ganze Gruppe zu verbessern (vgl. Elbing 1963, p. 32f.). Dies kann sie tun, indem sie Schülerinnen und Schüler, die sich in ungünstigen Positionen befinden und ausgeschlossen sind, durch Integrationsbemühungen, beispielsweise durch Berücksichtigung der Wahlwünsche und gezielte Förderung des Selbstbewusstseins, stützt. Weiter kann die Lehrperson vor allem bei der Zusammensetzung von Arbeitsgruppen regulierend eingreifen und dafür sorgen, dass die Gruppen bezüglich Leistungsfähigkeit und Ansehen möglichst heterogen sind, damit das soziale Ansehen der einzelnen Gruppen etwa gleich ist. Drittens kann die Gruppe durch die Thematisierung sozialer Beziehungen[82] und durch das Bewusstsein von Erwartungen, Vorurteilen, Ursachen und Zusammenhängen für Gruppenstrukturen lernen, eigene Spielregeln für das Zusammenleben in der Klasse aufzustellen (vgl. Petillon 1980, p. 134ff.).

[82] Petillon (1980) schlägt vor, «soziale Beziehungen« zum Unterrichtsthema zu machen, erläutert eine Unterrichtseinheit und empfiehlt einschlägige Literatur sowie Unterrichtsmaterialien» (Petillon 1980, p. 137; 141ff.).

Soziometrische Daten übernehmen somit eine diagnostische Aufgabe, indem sie auf die Notwendigkeit sowie die Möglichkeit der Veränderung der Gruppendynamik hinführen. Trotzdem dürfen die Ergebnisse nicht überbewertet werden und erlauben nur Aussagen, die sich auf das Kriterium, die Gruppe und einen begrenzten Zeitraum beziehen. Die Situation eines Individuums sieht ausserhalb des gewählten Kriteriums, vor allem aber ausserhalb der Gruppe, in der ein soziometrischer Test durchgeführt wurde, anders aus.

Zu guter Letzt: Die Lehrperson, die sich vorgenommen hat, an den sozialen Beziehungen ihrer Klasse etwas zu ändern, muss sich daher immer auch die Frage stellen, ob man den Schülerinnen und Schülern, denen man helfen will, tatsächlich einen Dienst erweist, wenn man ihre soziale Stellung thematisiert und offen legt. Es gehört zumindest zu den Aufgaben dieser Lehrperson, abzuschätzen, welche Konsequenzen eine Veränderung des sozialen Gefüges für einzelne Schülerinnen und Schüler mit sich bringt. Die Entscheidung zu helfen muss daher bewusst, gezielt und überlegt erfolgen.

4.6 Anwendungsmöglichkeiten der Beobachtung

Nach der Darstellung verschiedener Beobachtungssysteme im Abschnitt 4.5, sollen die Anwendungsmöglichkeiten derselben im Mittelpunkt der folgenden Betrachtung stehen. Dabei geht es einerseits um allgemeine Aspekte der Anwendungsmöglichkeiten (Abschnitt 4.6.1), die Sinn und Inhalt von Beobachtungen thematisieren und andererseits sollen Anwendungsansätze anhand des vorgestellten Beispiels (Abschnitt 4.5.1.4) aufgezeigt werden (Abschnitt 4.6.2).

4.6.1 Allgemeine Aspekte

Ein wissenschaftliches Beobachtungssystem kann für die praktizierende Lehrperson im Ablauf des täglichen Unterrichtens kaum in einem sinnvollen Aufwand-Ertrag-Verhältnis angewandt werden. Vorteile für Lehrpersonen ergeben sich aber, im Gegensatz zu schulpsychologischen Testsituationen, in den Möglichkeiten sorgfältiger Langzeitbeobachtungen in «natürlichen» Situationen des Schulalltags. Beobachtet werden können dabei nicht nur isolierte Situationen mit einzelnen Personen, sondern Gruppendynamiken in Klassenverbänden. Dies ermöglicht zum Beispiel auch wertvolle Hinweise im Bezug auf die sozial-emotionalen Entwicklung eines Kindes. Beobachtungen können von Lehrerinnen und Lehrern laufend eingesetzt werden und Informationen liefern, welche mit anderen Methoden nicht zugänglich wären.

Für die Lehrpersonen stellen sich in den Anwendungsmöglichkeiten grundsätzliche Fragen zur Beobachtung:

Wozu beobachten? Grund für eine gezielte Beobachtung könnte zum Beispiel ein Hinweis auf Mobbing in der Klasse oder ganz einfach eine Selbstreflexion der eigenen Unterrichtstätigkeit sein. Beobachtungen des Sozialverhaltens könnten dann die Grundlage für angemessene Beratung und gezielte Hilfe sein.

Was beobachten? In einem weiteren Schritt einer Untersuchung stellt sich die Frage: Was genau möchte ich wissen? Das Beobachtungsobjekt, der Gegenstand der Beobachtung und das Beobachtungssubjekt, die zu Beobachtenden, ergeben sich aus den Gründen der Beobachtung. Zum Beispiel wäre bei einer Observation zu Mobbing (Beobachtungsgegenstand) eine soziale Gruppe (Beobachtungssubjekt) notwendig, da sich Handlungen wie Schikanieren nur unter Mitbeteiligung anderer definieren und die isolierte Beobachtung einer Person zu keinen Resultaten führen würde.

Wie beobachten? Neben der Abgrenzung des Beobachtungszieles und des interessierenden Verhaltensbereiches betreffen Aspekte der Beobachtung die Methodik. Im Anschluss an eine theoretische Definition und Hypothesenbildung gilt es, das interessierende Phänomen zu operationalisieren, d.h.

einer Messung zugänglich zu machen. Je komplexer, mehrdimensionaler und abstrakter Beobachtungsziele sind, wie es zum Beispiel die Frage des Mobbings darstellt, desto erforderlicher wird eine Spezifikation des Konzepts, bei der die einzelnen Dimensionen des Begriffs herausgearbeitet werden müssen. Der Plan über den Ablauf der Beobachtung ergibt sich aus der Anzahl Beobachter, dem Beobachtungsobjekt, dem Zeitmanagement, den technischen Hilfsmitteln, etc. Während der Beobachtung ist zu berücksichtigen, dass die interessierenden Ausschnitte sich nicht klar von anderen Bereichen abgrenzen lassen und sich wechselseitig beeinflussen. Quantitativ erhobene Beobachtungen setzen in der Planung definierte Beobachtungsfelder voraus, welche im Laufe der Zeit gleich bleiben, qualitativ orientierte Methoden erlauben Modifikationen im Feld der Beobachtungsprozesse (vgl. Lamnek 2001, p. 271).

Wie auswerten? Die Auswertung der Datenerfassung beinhaltet eine Fehlerkontrolle und -bereinigung. Nicht selten sind zwei oder mehrere Pretests zur Prüfung der intersubjektiven Übereinstimmung erforderlich. Zeigen sich im Pretest Fehler, wird man das Erhebungsinstrument entsprechend korrigieren und anpassen müssen. An diesem Punkt stellt sich die Frage der Zweckmässigkeit bezüglich des Aufwands und der Realisierbarkeit. Der Umfang einer statistischen Analyse hängt von der Komplexität der Fragestellungen, der Art der Daten und auch vom Wissen der Lehrperson ab. Deskriptive Auswertungen erfordern in der Regel weniger Aufwand als multivariate Analysen und Hypothesentests (vgl. Diekmann 1999, p. 161 f.).

Die Ergebnisse sollten in einem Bericht festgehalten werden. Um eine sachgerechte Auseinandersetzung der Resultate und Fehlerkontrollen zu ermöglichen, sollten auch die methodischen Schritte in nachvollziehbarer Weise dokumentiert sein (vgl. Abbildung 4.6).

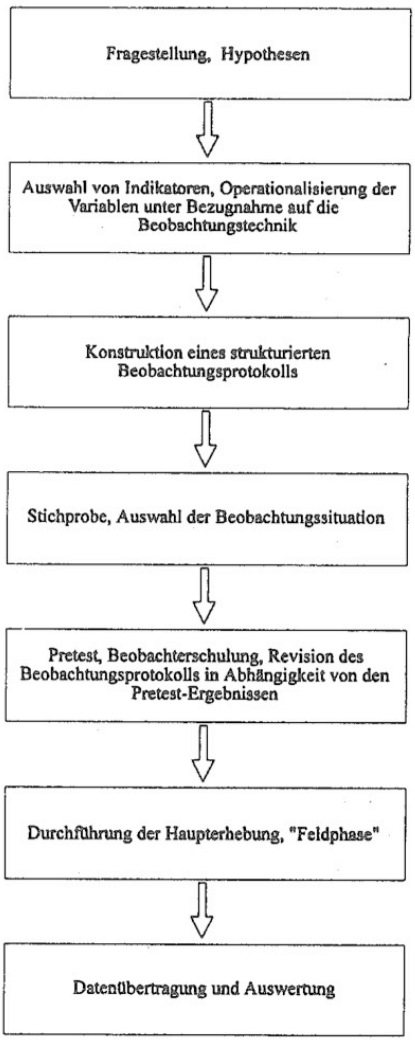

Abbildung 4.6: Ablauf einer Untersuchung mit strukturierter Beobachtungstechnik (aus: Diekmann 1999, p. 478)

Als letzter Schritt, welcher hier nicht erörtert wird, würde sich die Frage der Konsequenzen der Beobachtungsresultate stellen. Ergebnisse der Beobachtungen können von praktischem Interesse sein und die Untersuchung von Vorschlägen zur Lösung von Problemen mit einschliessen.

4.6.2 Anwendungsmöglichkeiten des dargestellten Beispiels

Die im Abschnitt 4.5.1.4 vorgestellte Beobachtungsskizze von Alsaker soll an dieser Stelle im Zusammenhang mit dem einleitenden Fallbeispiel nochmals aufgenommen werden. Im Vordergrund steht in erster Linie die Frage der Umsetzung in den Schulalltag. Aufgegriffen werden dabei auch die von Alsaker (2003) thematisierte Einzelbeobachtung bzw. der Einsatz eines Kategoriensystems.

Bezogen auf unser einleitendes Fallbeispiel könnte dies heissen, dass Herr Anderegg die Beobachtungsskizze als eine mögliche Methode bei seinen Beobachtungen einsetzt. Er müsste dabei in einem ersten Schritt die Beobachtungsinhalte, gekennzeichnet mit Beobachtung eins bis vier, festlegen. Mögliche Beobachtungsinhalte wären beispielsweise beleidigende respektive ausgrenzende Aussagen der erwähnten Knabengruppe gegenüber Kevin (vgl. Alsaker 2003, p. 71). Das zeitlich beschränkte Beobachten und visuelle Festhalten mehrerer solcher Interaktionen in Form einer Beobachtungsskizze, die in ihrer Gesamtheit das Phänomen des Mobbings fassbarer machen, führen zu einer systematischeren Erfassung der Mobbingsituation in der Klasse.

Eine andere Möglichkeit der systematischen Beobachtung, die Alsaker (ebd., p. 265) vorschlägt, ist die zeitlich beschränkte Einzelbeobachtung der Schülerinnen und Schüler. In einer solchen diskontinuierlichen Beobachtung könnte Anderegg verschiedene Lernende seiner Klasse zufällig oder bewusst auswählen und sie anhand seiner Beobachtungsinhalte für eine vorher festgelegte Zeit beobachten. Die Beobachtungsinhalte könnte Anderegg folgendermassen festlegen (ebd., p. 263):

- Ist Kevin systematisch Opfer von Gemeinheiten oder gibt es noch andere Opfer? Besteht die Täterschaft nur aus dieser Knabengruppe oder gibt es andere Schülerinnen und Schüler, die ebenfalls als Täter auftreten? Wie systematisch geht die Täterschaft beispielsweise bei Übergriffen auf das Eigentum von Kevin vor?
- Wie setzt sich Kevin in den verschiedenen Situationen zur Wehr? In welchem Verhältnis stehen die angreifenden Kinder zueinander, wie setzt sich diese Gruppe zusammen?
- Wie reagieren scheinbar nicht beteiligte Kinder auf die Mobbingsituationen? Wie reagiere ich als Lehrperson?
- Gibt es subtile Formen des Mobbings (schlechtes Reden, Ausschluss aus Gruppen)?

Die Protokollierung der Beobachtungen erfolgt schriftlich. Bei genügender Wiederholung kann der Klassenlehrer dadurch Rückschlüsse für das generelle Verhalten der Beobachteten ziehen.

Eine weitere hilfreiche Möglichkeit eröffnet sich gemäss Alsaker durch ein vorgängiges Zusammenstellen verschiedener zu erwartender Verhaltensarten nach Kategorien (ebd., p. 265). Anschliessend kann Beobachtetes durch Ankreuzen rasch protokolliert werden. Vorausgesetzt wird dabei jedoch, dass sich die Lehrperson ausführlich mit der Thematik des Mobbings auseinander gesetzt hat und dabei die wichtigen Indikatoren in den vorgefertigten Antwortmöglichkeiten berücksichtigt. Diese Beobachtungs- und Protokollierungsform eignet sich besonders bei Unterrichtsbeobachtungen durch Kolleginnen und Kollegen, stellt aber hohe Anforderungen an denjenigen, der die Beobachtungskategorien entwirft. Unten stehende Abbildung zeigt eine einfache Form eines Kategoriensystems zur Erhebung von Informationen grundsätzlicher Art.

Situation	Täter	Opfer	Lehrkraft anwesend	Wo	Art des Mobbing
			☐ ja ☐ nein	☐ Garderobe ☐ Pausenplatz ☐ Schulzimmer ☐ Schulweg ☐ Im Unterricht Fach:_____	☐ verbal ☐ physisch (Gewalt) ☐ gegenständlich ☐ isolieren

Abbildung 4.7: Unterrichtsbeobachtung Mobbing (in Anlehnung an Alsaker 2003, p. 300f.)

4.7 Grenzen der Unterrichtsbeobachtung

Die Komplexität sozialer Verhaltensweisen, Ereignisse und Situationen, die Beobachtungsgegenstand sind, und eine Fülle anspruchsvoller Anforderungen, die an die beobachtende Person gestellt werden, führen dazu, dass Beobachtungen zahlreiche Fehlerquellen und problematische Aspekte in sich bergen (vgl. Lamnek 2001, p. 275). Mit den Ausführungen in diesem Abschnitt werden zwei Absichten verfolgt: Sie sollen die Grenzen der Methode des Beobachtens verdeutlichen und zudem am Beispiel «Mobbing» die Grenzen der vorgestellten Möglichkeit für die Unterrichtsbeobachtung (Abschnitt 4.5.1.4) aufzeigen.

4.7.1 Allgemeine Grenzen der Beobachtungsmethode

Abbildung 4.8: Die Beobachtungsmethode und ihre Grenzen (aus: Sonntags Zeitung 30. 5. 2004, p. 119)

Die Beobachtung ist auf vielfältige Weise limitiert. Grenzen ergeben sich durch Raum und Zeit, der begrenzten Reichweite menschlicher Sinnesorgane, der gewählten Methodik des Beobachtungsinstruments, der eingesetzten Techniken, der Interpretation, der Urteilsfindung etc. Diekmann (1999, p. 458) nennt bei systematischen Beobachtungen vor allem zwei Problemkreise, welche zu beachten sind: Das Problem der Verzerrungen durch selektive Wahrnehmungen und das Problem der Fehlinterpretationen des beobachteten sozialen Geschehens.

Wahrnehmung ist Oberbegriff und Voraussetzung von Beobachtung. Beobachten charakterisiert die visuelle (und in der methodischen Beobachtung die akustische) Wahrnehmung in der Art, dass ein Ereignis bewusst zum Objekt der eigenen Aufmerksamkeit gemacht wird. Wahrnehmung ist aber kein Abbild der Aussenwelt, sondern ist verknüpft mit einem Akt der Konstruktion des Beobachters. Dabei ist nicht Selektivität an sich das Problem, sondern die Gefahr von Selektionsverzerrungen, d.h. die Werte der beobachtenden Variablen werden mit unterschiedlichen Wahrscheinlichkeiten registriert (ebd., p. 473). Selektioniert wird in dreifacher Hinsicht: «Es werden bestimmte Beobachtungsobjekte ausgewählt (Stichprobenselektion), es werden nur bestimmte Aspekte wahrgenommen (Wahrnehmungsselektion) und davon wiederum nur bestimmte Teile erinnert (Erinnerungsselektion)» (ebd., p. 44). Möglichkeiten, das Risiko einer Verzerrung durch selektive Wahrnehmung zu minimieren, sind detaillierte Instruktionen an die Beobachtenden, Verwendung strukturierter Beobachtungsprotokolle, Schulung von Beobachterinnen und Beobachtern, Einsatz mehrerer Beobachtenden und eventuell Verwendung technischer Hilfsmittel (vgl. ebd., p. 473).

Ein grundsätzliches Problem für die Beobachtungen im Unterricht stellt die Situation dar, dass die Lehrperson zugleich die Funktion des Beobachtenden und des Lehrenden einnimmt. Mit der institutionalisierten Alltagsrolle im sozialen Feld der Schule erhält die Lehrperson den Status einer teilnehmenden und aktiven Beobachterin. Die methodische Problematik liegt in der Konstellation einer nicht seltenen «Subjekt-Subjekt-Beziehung» von Lehrpersonen sowie Schülerinnen und Schülern. Die Lehrperson muss dabei im Sinne einer gerechten pädagogischen Diagnose die beobachtete Person «verobjektivieren». Die Gefahr der aktiven Teilnahme wird in der ethnologischen Forschung als das Problem des «going native» beschrieben, d.h. eine fehlende Distanz des Beobachters zu den zu beobachtenden Personen kann den Blick trüben, so dass der Wald vor lauter Bäumen nicht mehr gesehen wird.

Von wertfreier Beobachtung kann nie gesprochen werden. Bewertungserwartungen von Verhaltensweisen basieren auf sozialen und kulturellen Normen. Auswertungsberichte von Lern-, Arbeits- und Sozialverhalten beschreiben und bewerten in präskriptiven Sätzen. Aus einer Beobachtung

(«Daniel stört Chantal») wird auf eine Fähigkeit oder Persönlichkeitseigenschaft («Daniel ist aggressiv») geschlossen. Mechanismen und Effekte wie die «sich selbst erfüllende Prophezeiung» (self fulfilling prophecy) oder die «Tendenz zur Mitte» (keine extremen Urteile) beeinflussen dabei nicht nur den Prozess der Wahrnehmung, sie steuern und filtern auch die Interpretationen des Geschehens. Motive, Erfahrungen, Emotionen, Interessen, Ideologien und Vorstellungen usw. können in der Urteilsbildung einen erheblichen Einfluss ausüben (vgl. ebd., p. 61ff.). Wichtige Voraussetzungen zur Reduktion von Fehlerquellen sind neben der Anwendung systematischer Beobachtungen das Bewusstsein der eigenen Werthaltungen, Intoleranzen und Vorurteile. Und da noch kein Meister vom Himmel gefallen ist, bleibt die Übung wichtiger Bestandteil (vgl. Nuding 1997, p. 55 f.).

4.7.2 Grenzen der Unterrichtsbeobachtung am Beispiel Mobbing

Mobbing ist ein soziales Phänomen, ein komplexes Muster, das durch systematische Wiederholung gekennzeichnet ist und sich meist in subtiler Art und Weise abspielt. Erniedrigungen, Isolation und körperliche Angriffe finden oft nicht in der Gegenwart von Erwachsenen statt und sind dadurch nur bedingt sichtbar und beobachtbar. Unabhängig von den eingesetzten Beobachtungsmethoden begrenzt folglich bereits die Komplexität und die teilweise schlechte Beobachtbarkeit des Beobachtungsinhalts «Mobbing» die Beobachtungstätigkeit. Dieser Umstand lässt sich durch eine Untersuchung an Kindergärten zur Thematik Mobbing bestätigen (vgl. Alsaker 2003, p. 67ff.). Dreiviertel aller Mobbingsituationen finden beim freien Spielen, welches sich an verschiedenen Orten im Kindergarten oder auch draussen abspielt, statt, wo die Kontrolle von Erwachsenen eingeschränkt ist.

Die vorgestellte Beobachtungsskizze von Alsaker (vgl. Abschnitt 4.5.1.4) schafft die Möglichkeit einer Visualisierung, garantiert aber keinen Gewinn an Objektivität, Reliabilität und Validität. Sind die zu beobachtenden Indikatoren für Mobbing dem Beobachtenden nur teilweise bekannt, werden die Beobachtungsresultate trotz Skizze lückenhaft bleiben. Überdies greift die Beobachtungsskizze zu kurz, weil sie in erster Linie Übergriffe auf die Psyche und den Körper des Opfers und das Phänomen des «indirekten Mobbings»[83] darzustellen vermag. «Gegenständliches Mobbing»[84] (Alsaker 2003, p. 71) lässt sich indes weniger gut anhand einer Beobachtungsskizze darstellen.

Beobachtungsformen, die auf gut ausgearbeiteten Kategoriensystemen basieren, eignen sich besser für die Erhebung von Mobbingsituationen, da sie die verschiedenen Erscheinungsformen von Mobbing besser zu erfassen vermögen. Dabei ist jedoch in Rechnung zu stellen, dass Lehrpersonen oft nicht über die nötigen zeitlichen und teilweise auch fachlichen Ressourcen verfügen, um ein Kategoriensystem auszuarbeiten. Zudem können zu umfangreiche Beobachtungsvorhaben zur Überforderung der Lehrpersonen führen, weil sie als teilnehmende Beobachter agieren und parallel zur Beobachtungstätigkeit den Unterricht führen müssen. Folglich sollten die Beobachtung und Protokollierung derselben so einfach als möglich gestaltet werden, will man nicht auf technische Hilfsmittel zurückgreifen.

---

[83] Unter indirektem Mobbing wird der Ausschluss, die Isolation oder auch die Ausgrenzung des Opfers verstanden.

[84] Gegenständliches Mobbing richtet sich gegen das Eigentum des Opfers. Dazu gehören insbesondere das Verstecken, Stehlen oder auch Zerstören von Gegenständen, die dem Opfer gehören.

4.8 Fazit

Es gehört zur Grundaufgabe von Lehrpersonen, diagnostische Urteile sorgfältig und reflektiert vorzunehmen. Die dafür notwendigen Informationen können z.B. anhand einer systematisierten Unterrichtsbeobachtung erschlossen werden. Im Zusammenhang mit dem dargestellten Fallbeispiel wurde in diesem Kapitel zur Illustration soziales Verhalten fokussiert; wichtige Anwendungsfelder der Unterrichtsbeobachtung sind aber auch das Arbeits- und das Lernverhalten von Schülerinnen und Schülern. Auch wenn im schulischen Kontext nicht die gleichen Ansprüche betreffend Systematik wie in der Forschung gelten, sollte es Ziel der Lehrpersonen sein, ihre Beobachtungen mit den zur Verfügung stehenden Ressourcen zu systematisieren, um dadurch zu professionellen diagnostischen Urteilen zu gelangen. Dabei können auch relativ einfache Beobachtungssysteme, wie die Beobachtungsskizze von Alsaker (vgl. Abschnitt 4.5.1.4), gute Dienste leisten.

5 Verzeichnisse

5.1 Literaturverzeichnis

Adams, M. J. (1989). Thinking skills curricula: Their promise and progress. *Educational Psychologist, 24*, p. 25-77.

Alsaker, F. D. (2003). *Quälgeister und ihre Opfer. Mobbing unter Kindern – und wie man damit umgeht*. Bern: Verlag Hans Huber.

Alsaker, F. D. (2004). *Mobbing ist kein Kinderspiel [das Basis-Medienpaket]*. Bern: Schulverlag blmv.

Amelang, M.; Zielinski, W. (1997). *Psychologische Diagnostik und Intervention* (2., korrigierte, aktualisierte und überarbeitete Auflage). Berlin: Springer.

Amt für Berufsbildung und Berufsberatung des Kantons Solothurn. (2006). *Basic-Check*, [On-Line].Verfügbar unter: <http://www.so.ch/departemente/bildung-und-kultur/berufsbildung-mittel-und-hochschulen/zentrale-dienste/basic-check.html> [Februar 2012].

Arnold, K.-H. (1999). Diagnostische Kompetenz erwerben. Wie das Beurteilen zu lernen und zu lehren ist. *Pädagogik, 51*, p. 73-77.

Arnold, K.-H.; Jürgens, E. (2001). *Schülerbeurteilung ohne Zensuren*. Neuwied: Hermann Luchterhand.

Atteslander, P.; Cromm, J. (1995). *Methoden der empirischen Sozialforschung*. Berlin: Walter de Gruyter.

Aurin, K. (1991). *Gute Schulen – Worauf beruht ihre Wirksamkeit?* Bad Heilbrunn: Klinkhardt.

Basic-Check. (2004). *Basic-Check: Interpretations-Anleitung*.

BBT. (2004). *Lehrstellenbarometer: Lehrstellen etwas entspannter als im Vorjahr*. Bern: BBT.

Beck, K.; Krapp, A. (2001). Wissenschaftstheoretische Grundfragen der Pädagogischen Psychologie. In Krapp, A.; Weidenmann, B. (Eds.), *Pädagogische Psychologie. Ein Lehrbuch* (4. Aufl., p. 31-73). Weinheim: Beltz Psychologie Verlags Union.

Bibliographisches Institut; F.A. Brockhaus (2001). *Duden. Das Fremdwörterbuch*. Mannheim, Leipzig, Wien, Zürich: Dudenverlag.

Bloom, B. S. (1968). *Learning for Mastery*: University of California Evaluation Comment.

Borchert, J.; Knopf-Jerchov, H.; Dahbashi, A. (1991). *Testdiagnostische Verfahren in Vor-, Sonder- und Regelschulen. Ein kritisches Handbuch für Praktiker*. Heidelberg: Asanger.

Bortz, J. (1999). *Statistik für Sozialwissenschaftler* (4., vollständig überarbeitete Auflage). Berlin: Springer.

Coradi Vellacott, M.; Hollenweger, J.; Nicolet, M.; Wolter, S. C. (2003). *Soziale Integration und Leistungsförderung*. Neuchâtel: Bundesamt für Statistik (BFS), Schweizerische Konferenz der kantonalen Erziehungsdirektoren (EDK).

Criblez, L. (2002). Beurteilen als Grundlage von Promotion, Selektion und Berechtigung im Bildungssystem. In Rhyn, H. (Ed.), *Beurteilung macht Schule. Leistungsbeurteilung von Kindern, Lehrpersonen und Schule* (p. 84-93). Bern: Haupt.

Diekmann, A. (1999). *Empirische Sozialforschung. Grundlagen, Methoden, Anwendungen*. Reinbek bei Hamburg: Rowohlt Taschenbuch Verlag.

Dinges, E. (2002). *Systematische Beurteilung und Förderung schulischer Leistungen*. Horneburg: Persen.

Ditton, H. (1992). Profitieren Schüler vom Besuch 'guter Schulen'? Ergänzende Aspekte zu einer aktuellen Diskussion. In Siera, C.; Th, M. (Eds.), *Gesellschaft und Bildung* (p. 51-76). Pfaffenweiler: Centaurus.

Doyle, W. (1986). Classroom Organsization and Management. In Wittrock, M. C. (Ed.), *Handbook of Research on Teaching*. New York: Macmillan Publishing.

Dubs, R.; Delhees, K.; Metzger, C. (1973). *Leistungsmessung und Schülerbeurteilung*. Zürich: Schweizerischer Kaufmännischer Verein.

EDK. (2001). *Leitlinien der Schweizerischen Konferenz der kantonalen Erziehungsdirektoren EDK*, [On-Line]. Verfügbar unter: <http://edudoc.ch/record/24715/files/LL_EDK_d.pdf > [Februar 2012].

EDK. (2003a). Harmonisierung der obligatorischen Schule (HarmoS). Ein Projekt der EDK, [On-Line]. Verfügbar unter: <http://www.edudoc.ch/static/web/arbeiten/harmos/Kurz_info_d.pdf > [Februar 2012].

EDK. (2003b). *Europäisches Sprachenportfolio. Schweiz. Informationen und Hinweise für Menschen, die Sprachen lernen*. Bern: Schulverlag blmv AG.

Elbing, E. (1963). *Das Soziogramm der Schulklasse*. München, Basel: Ernst Reinhardt Verlag.

Ey-Ehlers, C. (2001). *Hochbegabte Kinder in der Grundschule – eine Herausforderung für die pädagogische Arbeit unter besonderer Berücksichtigung von Identifikation und Förderung*. Stuttgart: ibidem-Verlag.

Fassnacht, G. (1995). *Systematische Verhaltensbeobachtung. Eine Einführung in die Methodologie und Praxis*. München; Basel: Ernst Reinhardt Verlag.

Feger, B. (1988). *Hochbegabung – Chancen und Probleme*. Bern: Hans Huber.

Fend, H. (1981). *Theorie der Schule. (2. Aufl. ed.)*. München: Urban & Schwarzenberg.

Fisseni, H. J. (1997). *Lehrbuch der psychologischen Diagnostik (2., überarbeitete und erweiterte Auflage)*. Göttingen: Hogrefe.

Flammer, A. (1997). *Einführung in die Gesprächspsychologie*. Bern: Huber.

Gage, N. L. (1968). *Handbook of research on teaching. A project of the American Educational Research Association*. Chicago: Rand McNally.

Gage, N. L.; Berliner, D. C. (1996). *Pädagogische Psychologie* (5., vollständig überarbeitete Auflage). Weinheim: Psychologie Verlags Union.

Gardner, H. (1983). *Frames of mind*. New York: Basic Books.

Gardner, H. (1994). *Abschied vom IQ die Rahmen-Theorie der vielfachen Intelligenzen*. Stuttgart: Klett-Cotta.

Gerber, M. (2004). Auf dem Prüfstand der Swisscom. *Berner Schule,* p. 14-15.

Gipps, C. V. (1994). *Beyond testing towards a theory of educational assessment*. London: The Falmer Press.

Grissemann, H.; Weber, A. (1982). *Spezielle Rechenstörungen. Psychologische und kinderpsychiatrische Grundlagen der pädagogisch-therapeutischen Interventionen bei Kindern mit Dyskalkulie*. Bern: Hans Huber.

Groffmann, K. J. (1964). Die Entwicklung der Intelligenzmessung. In Heiss, R.; Michel, L.; Groffmann, K. J. (Eds.), *Handbuch der Psychologie* (Band 6: Psychologische Diagnostik, p. 147-199). Göttingen: Hogrefe.

Guilford, J. P. (1950). *Creativity*. In American Psychologist, 5, p. 444-454.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Haberlandt, K. (1999). *Human Memory*. Neddham Height: Viacom.

Häcker, T. (2003). Das Lernportfolio. Was die Theorie über die Bedeutung und den Einsatz von Lernportfolios sagen kann. *Forum Fortbildung, 4*, p. 10-12.

Hanses, P.; Rost, D. H. (1998). *Das «Drama» der hochbegabten Underachiever – «Gewöhnliche» oder «aussergewöhnliche» Underachiever?* In Zeitschrift für Pädagogische Psychologie, 12, p. 53–71.

Heinbokel, A. (1988). Hochbegabung. Erkennen, Probleme, Lösungswege. Baden-Baden: Nomos.

Helmke, A. (2003). *Unterrichtsqualität erfassen, bewerten, verbessern*. Seelze: Kallmeyer.

Helmke, A.; Hosenfeld, I.; Schrader, F.-W. (2004). *Vergleichsarbeiten als Instrument zur Verbesserung der Diagnosekompetenz von Lehrpersonen*. In: Arnold, Rolf; Griese, Christiane (Hrsg.): Schulleitung und Schulentwicklung. Voraussetzungen, Bedingungen, Erfahrungen. Baltmannsweiler: Schneider Verlag Hohengehren. p. 119-143.

Herrmann, J.; Höfer, C. (1999). *Evaluation in der Schule – Unterrichtsevaluation. Berichte und Materialien aus der Praxis*. Gütersloh: Verl. Bertelsmann-Stiftung.

Herzog, W.; Bärtschi, S.; Butters, A.; Hollenstein, A.; Retsch, S. et al. (2004). *Wirksamkeit von Schule und Unterricht. Modul APP SS 2004. Skript zur Veranstaltung «Wirksamkeit von Schule und Unterricht»*: Universität Bern, Virtueller Campus Pädagogik.

Hirsig, R. (2001). *Statistische Methoden in den Sozialwissenschaften: eine Einführung im Hinblick auf computergestützte Datenanalysen mit SPSS für Windows (Band 1; 3., erweiterte Auflage)*. Zürich: Seismo-Verlag.

Hogrefe, V. (1998). *Testkatalog 1998/99*. Göttingen: Hogrefe.

Holling, H.; Kanning, U. P.; Wittmann, A. J. (1999). *Hochbegabung: Forschungsergebnisse und Fördermöglichkeiten*. Göttingen: Hogrefe Verlag für Psychologie.

Hopp, A.; Lienert, G. (1965). *Eine Verteilungsanalyse von Gymnasialzensuren*. In Schule und Psychologie, 12, p. 139-150.

Hosenfeld, I.; Helmke, A.; Schrader, F.-W. (2002). Diagnostische Kompetenz: Unterrichts- und lernrelevante Schülermerkmale und deren Einschätzung durch Lehrpersonen in der Unterrichtsstudie SALVE. Zeitschrift für Pädagogik, 45, p. 65-82.

Huser, J. (1999). *Lichtblicke für helle Köpfe. Ein Wegweiser zur Erkennung und Förderung von hohen Fähigkeiten bei Kindern und Jugendlichen auf allen Schulstufen*. Zürich: Lehrmittelverlag des Kantons Zürich.

ILMES. *Internet-Lexikon der Methoden der empirischen Sozialforschung*, [On-Line]. Verfügbar unter: <http://www.lrz-muenchen.de/~wlm/ilmes.htm> [Februar 2012].

Ingenkamp, K. (1971). *Tests in der Schulpraxis eine Einführung in Aufgabenstellung, Beurteilung und Anwendung von Tests* (Bd. 16). Weinheim: Beltz.

Ingenkamp, K. (1977). *Schüler- und Lehrerbeurteilung: Empirische Untersuchungen zur pädagogischen Diagnostik* (Bd. 20). Weinheim: Beltz.

Ingenkamp, K. (1981). Verzeichnis der deutschsprachigen Schultests. Stand: Sommer 1980. In Jäger, R. S.; Ingenkamp, K.; Stark, G. (Eds.), *Tests und Trends 1981. Jahrbuch der Pädagogischen Diagnostik* (p. 147-163). Weinheim: Beltz.

Ingenkamp, K. (1995). *Die Fragwürdigkeit der Zensurengebung: Texte und Untersuchungsberichte*. Weinheim Basel: Beltz.

Ingenkamp, K. (1997). *Lehrbuch der pädagogischen Diagnostik*. Weinheim; Basel: Beltz Verlag.

Jachmann, M. (2003). *Noten oder Berichte? die schulische Beurteilungspraxis aus der Sicht von Schülern, Lehrern und Eltern* (Bd. 29). Opladen: Leske + Budrich.

Jackson, P. W. (1986). *The Practice of Teaching*. New York: Teachers College Press.

Jäger, R. S. (1992). Themen, Diskussionen und Trends. In Ingenkamp, K.; Jäger, R. S.; Petillon, H.; Wolf, B. (Eds.), *Empirische Pädagogik 1970-1990. Eine Bestandsaufnahme der Forschung in der Bundesrepublik Deutschland* (Bd. 1, p. 131-143). Weinheim: Deutscher Studien Verlag.

Jegge, J. (1976). *Dummheit ist lernbar*. Bern: Zytglogge.

Jones, L. V. (1971). The Nature of Measurement. In Thorndike, R. L. (Ed.), *Educational Measurement* (p. 335-355). Washington: American Council of Education.

Kastner, M. (1982). Zur Intelligenzmessung im Rahmen der Pädagogischen Diagnostik. In Klauer, K. J. (Ed.), *Handbuch der Pädagogischen Diagnostik (1. Band)* (p. 333-339). Düsseldorf: Pädagogischer Verlag Schwann.

Kleber, E. W. (1992). *Diagnostik in pädagogischen Handlungsfeldern. Einführung in Bewertung, Beurteilung, Diagnose und Evaluation*. Weinheim: Juventa.

Kluckhohn, C.; Murray, H. A. (1948). *Personality Formation: The Determinants* (4. Ausgabe, 1950). New York: Alfred A. Knopf.

Knellwolf, T.; Vögeli-Mantovani, U. (2003). Wenn die Schulnoten nichts mehr gelten. *Bildung Schweiz, 12*, p. 19.

Köhne, H. (1979). Verhaltensdiagnostische Methoden. In Köhne, H.; Klippstein, E. (Eds.), *Pädagogische Verhaltensdiagnostik in der Praxis. Nichtinstrumentelle Untersuchungsverfahren für Sozialpädagogen, Erzieher und Lehrer*. Freiburg, Basel, Wien: Herder.

Konradt, H.-J. (1975). *Lehrziele, Schulleistung und Leistungsbeurteilung*. Düsseldorf: Pädagogischer Verlag Schwann.

Krapp, A. (1994). Diagnose und Prognose. In Weidenmann, B.; Krapp, A. (Eds.), *Pädagogische Psychologie. 3. Auflage* (p. 567-630). Weinheim: Beltz.

Kriz, J. (1973). *Statistik in den Sozialwissenschaften*. Reinbek: Rowohlt.

Kronig, W.; Haeberlin, U.; Eckhart, M. (2000). *Immigrantenkinder und schulische Selektion. Pädagogische Visionen, theoretische Erklärungen und empirische Untersuchungen zur Wirkung integrierender und separierender Schulformen in den Grundschuljahren*. Bern: Verlag Paul Haupt.

Kunz-Heim, D. (2002). *Qualität durch Qualifizierung. Lehrerbeurteilung als Instrument zur Förderung von Qualität im Unterricht*. Weinheim; München: Juventa.

Lamnek, S. (2001). Beobachtung. In Hug, T. (Ed.), *Wie kommt Wissenschaft zu Wissen? Einführung in die Forschungsmethodik und Forschungspraxis*. Hohengehren: Schneider.

Lamprecht, M.; Stamm, H. (1996). *Soziale Ungleichheit im Bildungswesen. (Eidgenössische Volkszählung 1990, hrsg. vom Bundesamt für Statistik BFS)*. Bern: BFS.

Landert Farago Davatz & Partner (Hrsg.) (2002). *Klassencockpit – Evaluation*. Verfügbar unter: <http://edudoc.ch/record/3003/files/zu03038.pdf> [Februar 2012]

Ledl, V. (1997). *Kinder beobachten und fördern eine Handreichung zur gezielten Beobachtung und Förderung von Kindern mit besonderen Lern- und Erziehungsbedürfnissen bzw. sonderpädagogischem Förderbedarf*. Wien: Verlag Jugend und Volk.

Lienert, G. A. (1969). *Testaufbau und Testanalyse* (3. Auflage). Weinheim: Beltz.

Lienert, G. A. (1991). *Testaufbau und Testanalyse* (5. Auflage). Weinheim: Beltz.

Lienert, G. A.; Raatz, U. (1994). *Testaufbau und Testanalyse (5., völlig neubearbeitete und erweiterte Auflage)*. Weinheim: Beltz.

Lissmann, U.; Paetzold, B. (1984). Zur Effektivität von Schülerselbstkorrektur und häufiger Leistungsrückmeldung. Eine empirische Untersuchung. In *Zeitschrift für Pädagogik*, 30 (6), p. 817-833.

Lissmann, U. (2001). Die Schule braucht eine neue Pädagogische Diagnostik. Formen, Bedingungen und Möglichkeiten der Portfoliobeurteilung. *Die Deutsche Schule, 93* (4), p. 486-497.

Lukesch, H. (1998). *Einführung in die pädagogisch-psychologische Diagnostik*. Regensburg: Roderer.

Martin, E.; Wawrinowski, U. (1991). *Beobachtungslehre Theorie und Praxis reflektierter Beobachtung und Beurteilung*. Weinheim: Juventa Verlag.

Mattl, W. (1978). *Notengeben in der Schule*. Freiburg: Herder.

Medley, D. M.; Impellitterie, J. T.; Smith, L. H. (1966). *Coding Teachers' Verbal Behavior in the Classroom. A manual for users of OScAR 4V*. New York: University.

Messner, H. (1978). *Wissen und Anwenden. Zur Psychologie im Unterricht*. Stuttgart: Klett-Cotta.

Mollenhauer, K.; Rittelmeyer, C. (1977). *Methoden der Erziehungswissenschaft* (Bd. 3). München: Juventa.

Moser, U. (2003). *Klassencockpit im Kanton Zürich. Ergebnisse einer Befragung von Lehrerinnen und Lehrern der 6. Klassen über ihre Erfahrungen im Rahmen der Erprobung von Klassencockpit im Schuljahr 2002/03. Bericht zuhanden der Bildungsdirektion des Kantons Zürich*, [On-Line]. Verfügbar unter: <http://edudoc.ch/record/3012/files/zu03064.pdf > [Februar 2012].

Moser, U.; Rhyn, H. (2000). *Lernerfolg in der Primarschule eine Evaluation der Leistungen am Ende der Primarschule* (Bd. 28). Aarau: Bildung Sauerländer.

Neuenschwander, M. P.; Hirt, U.; Balmer, T.; Gasser, A.; Goltz, S. et al. (2003). *Eltern, Lehrpersonen und Schülerleistungen. Dokumentation des Leistungstests Mathematik*.

Nuding, A. (1997). *Beurteilen durch Beobachten. Pädagogische Diagnostik im Schulalltag*. Baltmannsweiler: Schneider Verlag Hohengehren.

Oelkers, J. (2002). *Leistungen und Noten: Probleme der Schülerbeurteilung*, [On-Line]. Verfügbar unter: <http://edudoc.ch/record/29498/files/14.pdf > [Februar 2012].

Orth, B. (1974). *Einführung in die Theorie des Messens*. Stuttgart: Kohlhammer.

Petillon, H. (1980). *Soziale Beziehungen in Schulklassen*. Weinheim, Basel: Beltz Verlag.

Plössnig, S. (1997). Alternative Leistungsbeurteilung – eine Folge von differenzierendem und individualisierendem Unterricht in der Grundstufe 1. *Erziehung und Unterricht, 147* (5-6), p. 663-667.

Plotke, H. (2003). *Schweizerisches Schulrecht*. Bern: Haupt.

Rheinberg, F. (1980). *Leistungsbewertung und Lernmotivation*. Göttingen: Hogrefe.

Rheinberg, F. (1998). Bezugsnormorientierung. In Rost, D. H. (Ed.), *Handwörterbuch Pädagogische Psychologie* (p. 39-43). Weinheim: Beltz.

Rheinberg, F.; Bromme, R.; Minsel, B.; Winteler, A.; Weidemann, B. (2001). Die Erziehenden und Lernenden. In Krapp, A.; Weidemann, B. (Eds.), *Pädagogische Psychologie* (4., vollständig überarbeitete Auflage, p. 271-355). Weinheim: Beltz.

Rhyn, H. (2002). *Beurteilung macht Schule. Leistungsbeurteilung von Kindern, Lehrpersonen und Schule*. Bern: Haupt.

Rhyn, H.; Moser, U. (2002). Schülerinnen und Schüler beurteilen. In Rhyn, H. (Ed.), *Beurteilung macht Schule. Leistungsbeurteilung von Kindern, Lehrpersonen und Schule* (p. 25-34). Bern: Haupt.

Richter, A. (2003). Sich selbst beurteilen lernen – das Portfolio als alternatives Bewertungsinstrument. *Der fremdsprachliche Unterricht Englisch, 6*, p. 44-47.

Rost, D. (1989). Zum Thema «Hochbegabung». In Oberkreisdirektor des Kreises Neuss (Ed.), *Bundesmodell Begabtenförderung im Kreis Neuss*. Neuss.

Rost, D. H. (1998). Hochbegabung. In Rost, D. H. (Ed.), *Handwörterbuch Pädagogische Psychologie* (p. 177-182). Weinheim: Beltz.

Roth, L. (1978). Empirische Forschungsmethoden. In Roth, L. (Ed.), *Methoden erziehungswissenschaftlicher Forschung*. Stuttgart: Kohlhammer.

Sacher, W. (2001). *Leistungen entwickeln, überprüfen und beurteilen. Grundlagen, Hilfen und Denkanstösse für alle Schularten*. Bad Heilbrunn/Obb: Klinkhardt.

Schilling, S.; Prochinig, T. (1998). *Dyskalkulie, Rechenschwäche*. Schaffhausen: Schubi-Verlag.

Schmid, R. (1977). *Intelligenz- und Leistungsmessung: Geschichte und Funktion psychologischer Tests*. Frankfurt: Campus.

Schrader, F.-W. (1989). *Diagnostische Kompetenzen von Lehrern und ihre Bedeutung für die Gestaltung und Effektivität des Unterrichts*. Frankfurt a.M.: Lang.

Schrader, F.-W. (2001). Diagnostische Kompetenz von Eltern und Lehrern. In Rost, D. H. (Ed.), *Handwörterbuch Pädagogische Psychologie* (p. 91-96). Weinheim: Psychologie Verlags Union.

Schröter, G. (1971). *Die ungerechte Aufsatzzensur*. Bochum.

Seiffge-Krenke, Inge (1974). *Probleme und Ergebnisse der Kreativitätsforschung*. Bern; Stuttgart; Wien: Hans Huber.

Stamm, M. (1999). Einführung in die Thematik. In Schweizerische Koordinationsstelle für Bildungsforschung SKBF (Ed.), *Begabungsförderung in der Volksschule. Umgang mit Heterogenität. Trendbericht Nr.2*. (p. 10-28). Aarau.

Stamm, M. (1992): *Hochbegabungsförderung in den Deutschschweizer Volksschulen*. Historische Entwicklung, Zustandsanalyse, Entwicklungsplan. Dissertation Universität Zürich.

Steiner, G. (2001). Lernen und Wissenserwerb. In Krapp, A.; Weidemann, B. (Eds.), *Pädagogische Psychologie* (p. 137-205). Weinheim: Beltz.

Steltzer, S. (1998). *Wenn die Wörter tanzen. Legasthenie und Schule*. München: Ariston.

Strittmatter, A. (2004). Noten sind praktisch – und unprofessionell. Elementarwissen zur Noten-Kontroverse – und die (bekannte) professionelle Alternative. *Bildung Schweiz, 6*.

Süllwold, F. (1983). Pädagogische Diagnostik. In Groffmann, K. J.; Michel, L. (Eds.), *Intelligenz- und Leistungsdiagnostik* (Band II der Enzyklopädie für Psychologie, p. 307-386). Göttingen: Verlag für Psychologie.

Sydow von, K. (2004). *Beobachtung*, [On-Line]. Universität Duisburg. Verfügbar unter: <http://www.uni-duisburg.de/FB2/PS/PER/vonSydow/WS0304/v-diagn05-beobachtung.pdf> [Januar 2011].

Tent, L.; Stelzl, I. (1993). *Pädagogisch-psychologische Diagnostik. Band 1: Theoretische und methodische Grundlagen* (1). Göttingen: Hogrefe.

Tewes, U.; Rossmann, P.; Schallberger, U. (2000a). HAWIK-III, Handbuch und Testanweisung (3. Auflage). Bern: Verlag Hans Huber.

Tewes, U.; Rossmann, P.; Schallberger, U. (2000b). *Hamburg-Wechsler-Intelligenztest für Kinder (HAWIK-III) (*Testkoffer*)*. Bern: Verlag Hans Huber.

Ulshöfer, R. (1963). *Gymnasium und Sozialwissenschaften: Wege zur Demokratisierung der Schule*. Heidelberg: Quelle & Meyer.

Urban, D. (1982). Systematische Statistik für die computergestützte Datenanalyse. Ein Handbuch zum Programm- Paket SYSTAT. Stuttgart: Gustav Fischer.

Valkanover, S.; Alsaker, F. D.; Svrcek, A.; Kauer, M. (2004). *Mobbing ist kein Kinderspiel. Arbeitsheft zur Prävention in Kindergarten und Schule*. Bern: Schulverlag.

Vögeli-Mantovani, U. (1998). Lernberichte, Beurteilungsgespräche und Zeugnisse. Zur Form der Berichterstattung in den Kantonen. *Schweizer Schule, 5*, p. 30-37.

Vögeli-Mantovani, U. (1999a). Beurteilung in der Schule. Mehr Berichte und Gespräche, weniger Notenzeugnisse. *LCH-Aktuell, 18.11.1999*.

Vögeli-Mantovani, U. (1999b). *Mehr fördern, weniger auslesen. Zur Entwicklung der schulischen Beurteilung in der Schweiz* (Nr. 3). Aarau: Schweizerische Koordinationsstelle für Bildungsforschung.

Vögeli-Mantovani, U. (2001). Das Kreuz mit den Noten. *Bildung Schweiz, 10*, p. 21.

Wahl, D.; Weinert, F. E.; Huber, G. L. (1997). *Psychologie für die Schulpraxis ein handlungsorientiertes Lehrbuch für Lehrer* (6. Auflage). München: Kösel-Verlag.

Weinert, F. E.; Lingelbach, H. C. (1995). Teaching expertise: Theoretical conceptualizations, empirical findings, and some consequences for teacher training. In Hoz, R.; Silberstein, M. (Eds.), *Partnerships of schools and institutions of higher education in teacher development*. (p. 293-302). Beer-Shera, Israel: Ben Gurion University of the Negev Press.

Weinert, F. E.; Schrader, F.-W.; Helmke, A. (1990). Educational expertise: Closing the gap between educational research and classroom practice. *School Psychology international, 11*, p. 163-180.

Weiss, R. (1965). Über die Zuverlässigkeit der Ziffernbenotung bei Ausätzen. *Schule und Psychologie, 9*, p. 257-269.

Wild, K.-P.; Hofer, M.; Pekrun, R. (2001). Psychologie des Lerners. In Krapp, A.; Weidemann, B. (Eds.), *Pädagogische Psychologie; Ein Lehrbuch* (p. 207-270). Weinheim: Beltz.

Wild, K.-P.; Krapp, A. (2001). Pädagogisch-psychologische Diagnostik. In Krapp, A.; Weidemann, B. (Eds.), *Pädagogische Psychologie* (4., vollständig überarbeitete Auflage, p. 513-563). Weinheim: Beltz.

Winner, E. (1998). *Hochbegabt*. Stuttgart: Klett-Cotta.

Ziefuss, H.; Petrat, G.; Roth, L. (1978). *Methoden der Unterrichtsbeobachtung*. Braunschweig: Westermann.

Zoppi, S. (2003). Klassencockpit – ein Instrument zur Qualitätssicherung und -entwicklung? Möglichkeiten und Grenzen des Transfers eines Instruments auf die Sekundarstufe II.

5.2 Internetseiten

ADD-Online – ADHS ADS Hyperaktivitätsstörung: www.adhs.ch, [Februar, 2012].

Basic-Check: www.basic-check.ch, [Februar, 2012].

ILMES – Internet-Lexikon der Methoden der empirischen Sozialforschung: www.lrz-muenchen.de/~wlm/ilmes.htm, [Februar, 2012].

Klassencockpit: www.klassencockpit.ch, [Februar, 2012].

Mathematische Basteleien (Der Turm von Hanoi): www.mathematische-basteleien.de/hanoi.htm, [Februar, 2012].

Multicheck: www.multicheck.ch, [Februar, 2012].

PISA: www.pisa.admin.ch, [Februar, 2012].

Psychologie Information ZPID: www.zpid.de, [Februar, 2012].

Schweizerische Konferenz der kantonalen Erziehungsdirektoren: www.edk.ch, [Februar, 2012].

Sprachenportfolio: www.sprachenportfolio.ch, [Februar, 2012].

Testzentrale der Schweizer Psychologen: www.testzentrale.ch, [Februar, 2012].

TIMSS: www.timss.mpg.de, [Februar, 2012].