



Florian Hofmann

Selbsteinschätzungen im Aufsatz- und Schreibunterricht

Eine empirische Untersuchung zu den Effekten von
formativen Leistungsdiagnosen und Selbsteinschätzungen
auf die Schreibkompetenz

Hofmann

Selbsteinschätzungen im Aufsatz- und Schreibunterricht

Florian Hofmann

Selbsteinschätzungen im Aufsatz- und Schreibunterricht

Eine empirische Untersuchung zu den Effekten
von formativen Leistungsdiagnosen und
Selbsteinschätzungen auf die Schreibkompetenz

Verlag Julius Klinkhardt
Bad Heilbrunn • 2017

k

Ich danke meiner Frau Julia für ihre bedingungslose Unterstützung.

Die vorliegende Arbeit wurde von der Philosophischen Fakultät und Fachbereich Theologie der Friedrich-Alexander-Universität Erlangen-Nürnberg unter dem Titel „Die Entwicklung und Evaluation eines domänenspezifischen Selbsteinschätzungsinstruments als Element formativer Leistungsmessung in der Sekundarstufe I“ als Dissertation angenommen.

Gutachter: Prof. Dr. Uwe Maier, Prof. Dr. Thomas Eberle

Tag der mündlichen Prüfung: 21.06.2016

Dieser Titel wurde in das Programm des Verlages mittels eines Peer-Review-Verfahrens aufgenommen.
Für weitere Informationen siehe www.klinkhardt.de.

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet abrufbar über <http://dnb.d-nb.de>.

2017.kg © by Julius Klinkhardt.

Das Werk ist einschließlich aller seiner Teile urheberrechtlich geschützt.

Jede Verwertung außerhalb der engen Grenzen des Urheberrechtsgesetzes ist ohne Zustimmung des Verlages unzulässig und strafbar. Das gilt insbesondere für Vervielfältigungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung und Verarbeitung in elektronischen Systemen.

Coverabbildung: © Goldfaery / istockphoto.

Satz: Kay Fretwurst, Spreewald.

Druck und Bindung: AZ Druck und Datentechnik, Kempten.

Printed in Germany 2017.

Gedruckt auf chlorfrei gebleichtem alterungsbeständigem Papier.

ISBN 978-3-7815-2179-7

Inhalt

Danksagung und persönliche Gedanken zur Arbeit	7
Zusammenfassung	9
1 Einleitung	11
2 Theoretische Grundlagen	13
2.1 Formative Leistungsmessungen	13
2.1.1 Einführung in die Thematik	13
2.1.2 Formative Leistungsmessung: Annäherung, Definition, Erläuterung und Abgrenzung	14
2.2 Der Erhebungsgegenstand: Die Kernkompetenz „Inhalte zusammenfassen“	39
2.2.1 Schreiben als Schlüsselqualifikation des Deutschunterrichts	39
2.2.2 Schreibaufgaben als Sonderfall offener Aufgabenformate	40
2.2.3 Die Beurteilung und Bewertung von Schreibaufgaben	41
2.2.4 Rückmeldungen (Feedback) zu Schreibaufgaben	43
2.2.5 Die Bedeutung der Kernkompetenz „Inhalte zusammenfassen“	59
2.2.6 Definition und Erläuterung der Kernkompetenz „Inhalte zusammenfassen“	61
2.3 Selbsteinschätzungen in schulischen Lehr-Lernprozessen	63
2.3.1 Selbsteinschätzen als Konstituente „selbständigen Lernens“	63
2.3.2 Selbsteinschätzungen im Unterricht	66
2.3.3 Empirische Befunde zu „Selbsteinschätzungen“	96
3 Forschungsdesiderat und eigene Fragestellungen	101
3.1 Zusammenfassung der Befundlage, Literaturüberblick und Forschungsdesiderat	101
3.2 Eigene Fragestellungen und Hypothesen	105
3.2.1 Forschungsfrage F1: Besitzen formative Kompetenzausprägungen bei ganzheitlichen Leistungen einen positiven Effekt auf die Leistung von Schülerinnen und Schülern bei Schreibaufgaben?	105
3.2.2 Forschungsfrage F2: Besitzen Selbsteinschätzungen im Rahmen formativer Leistungsmessungen bei ganzheitlichen Leistungen einen positiven Effekt auf die Kompetenzausprägung von Schülerinnen und Schülern bei Schreibaufgaben?	106
3.2.3 Forschungsfrage F3: Nähern sich die Selbsteinschätzungen der Schülerinnen und Schüler zur ganzheitlichen Leistung „Inhalte zusammenfassen“ durch Übung und Feedback dem Urteil des Lehrenden an?	107
4 Forschungsmethode	109
4.1 Überblick über das Design der Gesamtuntersuchung	109
4.2 Untersuchungsablauf der Gesamtuntersuchung	110

4.3	Untersuchungsdesign	113
4.3.1	Erhebungsmethode	114
4.3.2	Zeitliche Dimension der Erhebung	114
4.3.3	Untersuchungspläne zu den einzelnen Forschungsfragen	115
4.3.4	Testverfälschung und Kontrolle der Störvariablen	124
4.4	Stichprobe	136
4.4.1	Stichprobenkonstruktion	136
4.4.2	Stichprobenbeschreibung	137
4.4.3	Analyse von und Umgang mit fehlenden Werten	138
4.5	Messinstrument und Materialien	143
4.5.1	Messinstrument: Kriterienraster sowie Selbsteinschätzungs- und Beurteilungsbogen	143
4.5.2	Materialien	163
4.6	Gütekriterien	172
4.6.1	Objektivität	172
4.6.2	Reliabilität	173
4.6.3	Validität	174
5	Ergebnisse	179
5.1	Anmerkungen zum Ergebnisteil	179
5.2	Hypothesenprüfung	180
5.2.1	Hypothese 1	180
5.2.2	Hypothese 2	198
5.2.3	Hypothese 3	206
6	Diskussion	227
6.1	Forschungsfrage 1: Besitzen formative Leistungsmessungen bei ganzheitlichen Leistungen einen positiven Effekt auf die Kompetenzausprägung von Schülerinnen und Schülern bei Schreibaufgaben?	227
6.2	Forschungsfrage 2: Besitzen Selbsteinschätzungen im Rahmen formativer Leistungsmessungen bei ganzheitlichen Leistungen einen positiven Effekt auf die Kompetenzausprägung von Schülerinnen und Schülern bei Schreibaufgaben?	233
6.3	Forschungsfrage 3: Nähern sich die Selbsteinschätzungen der Schülerinnen und Schüler [...] dem Urteil des Lehrenden an?	234
6.4	Design- und Methodenkritik	236
7	Ausblick	243
	Verzeichnisse	245
	Literaturverzeichnis	245
	Abbildungsverzeichnis	267
	Tabellenverzeichnis	269

Danksagung und persönliche Gedanken zur Arbeit

Um ein Projekt nach über vier Jahren abzuschließen und eine Qualifikationsarbeit im vorliegenden Umfang zu verfassen, bedarf es neben eigenem Engagement und Durchhaltevermögen vor allem einer Vielzahl an Personen, die einem unterstützend und begleitend zur Seite stehen. Zurückblickend hatte ich das große Glück, in den unterschiedlichen Phasen der Entstehung jeweils auf Menschen zu treffen, die mich der Fertigstellung auf gänzlich verschiedene Weise ein Stück näher brachten – jenen sei dieser erste Part der Arbeit in Dankbarkeit gewidmet.

Leider wird es mir schon aufgrund der vielen verschiedenen Etappen des Entstehungsprozesses nicht gelingen, alle Personen, die mich während der fast fünfjährigen Arbeitsphase begleitet haben, namentlich zu erwähnen. Ich bitte dies ausdrücklich zu entschuldigen; ihr seid möglicherweise nicht genannt, ich habe eure Hilfe aber nicht vergessen – seid euch dessen sicher.

Zu Beginn möchte fast selbstredend Herrn Prof. Dr. Uwe Maier danken, der mir eine Stelle als Mitarbeiter am Lehrstuhl für Schulpädagogik der FAU Erlangen-Nürnberg offerierte und als Erstgutachter diese Arbeit betreute. Ich danke ihm für seine Unterstützung, besonders für seine schnellen, konstruktiven und motivierenden Rückmeldungen. Während meiner gesamten Schaffensphase hegte er nie Zweifel am Gelingen des Projekts, zeigte sich stets interessiert und hatte immer ein offenes Ohr für meine Belange. Ihm habe ich es zu verdanken, dass ich Interesse am wissenschaftlichen Arbeiten fand und mein erster Arbeitsabschnitt von großer Motivation und enormen Engagement geprägt war.

Herrn Prof. Dr. Thomas Eberle gilt ebenso mein Dank, er hat sich ohne große Bedenkzeit nahezu spontan bereit, erklärt als Zweitgutachter zu fungieren. Im Laufe der Entstehung durfte ich ihm mehrfach längere Textstellen zur Lektüre schicken; sogar während einer Schiffsreise, welche er im Rahmen seines Projektes „Klassenzimmer unter Segeln“ unternahm, begleiteten ihn Teile meiner Arbeit.

Während der ersten Entwicklungsphase war mir auch Frau Dr. Sigrid Zeitler eine große Hilfe. Stets interessierte sie sich für mein Projekt und stand als konstruktive Gesprächs- und Diskussionspartnerin parat. Von ihrem reichen Wissen und ihrer großen Begeisterung für die Erziehungswissenschaft durfte ich während ihrer Zeit als wissenschaftliche Mitarbeiterin am Lehrstuhl für Schulpädagogik oftmals profitieren. Danke Sigrid! Auch Frau Dr. Eva Fritzsche gebührt mein Dank. Mehrfach beriet sie mich bei methodischen Überlegungen, geduldig widmete sie sich allen Rückfragen, gleich wie banal sie aus heutiger Sicht auch gewesen sein mögen. Während der Datenerhebungsphase und der ersten Auswertung unterstützte mich am Lehrstuhl vor allem Herr Dr. Nikolaus Schröck. Er bemühte sich während seiner Zeit als Lehrstuhlvertretung um eine Verlängerung meiner Abordnung und sorgte nicht zuletzt durch regelmäßige Forschungskolloquien für eine permanente Weiterentwicklung meiner Arbeit. Über zahlreiche Fachgespräche entwickelte sich zudem auch ein persönliches Verhältnis, für welches ich bis heute dankbar bin.

Selbstredend dürfen mit Blick auf den der Arbeit zugrunde liegenden Datensatz die vielen Schülerinnen und Schüler sowie Lehrkräfte nebst Schulleitung nicht vergessen werden: Ohne deren Bereitschaft, an diesem sehr aufwändigen Projekt teilzunehmen, wäre es nicht zu einer Umsetzung gekommen. Die Teilnehmerinnen und Teilnehmer hatten über mehrere Wochen immer wieder zusätzliche Schreibaufgaben zu erledigen, die Lehrkräfte mussten diese neben ihren Alltagspflichten korrigieren – beides nicht hoch genug einzuschätzende Leistungen, danke dafür. Euer Einsatz und eure nach und nach eintreffenden Ergebnisse lösten einen zweiten

Motivationsschub bei mir aus, der anfallende Schreibblockaden und partiell auftauchende Resignation vor der zu erbringenden Gesamtleistung entgegenwirkte.

Meiner augenblicklichen Vorgesetzten am Lehrstuhl für Schulpädagogik, Frau Prof. Dr. Michaela Gläser-Zikuda, danke ich für ihr permanentes Interesse an meinem Projekt und die Wertschätzung meiner Arbeit. Ihr Vertrauen in meine Fähigkeiten half mir vor allem in den letzten schweren Monaten und Wochen der Fertigstellung.

Insgesamt muss ich wohl vor allem die Personen in den Mittelpunkt meiner Danksagung stellen, die mich in der Schlussphase der Dissertation begleiteten. So soll in diesem Zusammenhang nicht unerwähnt bleiben, dass zeitweise aus Freude am Forschen und Enthusiasmus für eigenes wissenschaftliches Arbeiten Gefühle der Resignation erwuchsen. Über Zweifel an der eigenen Leistungsfähigkeit wurde das Dissertationsprojekt mehr und mehr zur Belastung, mehrfach spielte ich mehr oder weniger ernsthaft mit dem Gedanken, die Arbeit nicht zu beenden. Allen Kolleginnen und Kollegen sowie Freundinnen und Freunden, die mich immer wieder aus einem (Forschungs)Tal holten, gebührt besonderer Dank.

In erster Linie ist hier meine geschätzte Kollegin, Frau Dr. Nicole Wolf, zu nennen. Ihre aufmunternden Worte und ihre Ermahnungen, das Abschließen der Arbeit nicht aus dem Auge zu verlieren, stellten einen bedeutsamen Motivationsfaktor dar. Auch Frau Susi Klaß möchte ich danken, als „Leidensgenossin“ konnte sie meine Gedanken stets nachvollziehen. Gleiches gilt für Herrn Dr. Klaus Wild, mit dem mich nicht nur meine Liebe zu einem großen Nürnberger Fußballverein verbindet.

Außerhalb des beruflichen Umfeldes bedanke ich mich bei allen Freundinnen und Freunden, denen ich während der Dissertationsphase zu wenig Aufmerksamkeit und Zeit zu teil werden ließ.

Zum Schluss danke ich denen, die mich nicht nur während eines Lebensabschnittes begleiten, sondern die mein Leben sind: Meiner Familie. Ich bedanke mich bei meinem Vater Rainer Hofmann, der mich lehrte, dass vieles erreichbar ist, auch wenn es manchmal einiger Umwege bedarf. Ich danke meiner Mutter Elisabetha Hofmann, die stets so ohne Zweifel an mich glaubte, wie es wohl nur eine Mutter vermag. Vor allem aber danke ich meiner geliebten Frau Julia, die mich während aller Phasen auf jegliche nur erdenkliche Weise unterstützte und mich oftmals entbehren musste. Sie fungierte als Organisator, Kritiker, Motivator und Korrektor; viele Male musste sie sich der Thematik meiner Arbeit im Gespräch widmen, oftmals war sie meine wichtigste Diskussionspartnerin. Und ich danke meinen Söhnen Leopold Maximilian und Luis Nicolas; dafür, dass sie auf unterschiedliche Weise dafür sorgen, dass man bei aller Fokussierung auf das Dissertationsprojekt nicht aus dem Auge verliert, was wirklich zählt im Leben; danke, dass es euch gibt und dafür, dass ihr so seid, wie ihr seid.

Zusammenfassung

Es gilt mittlerweile als empirisch gesichert, dass formative Leistungsmessungen große Effekte auf die Schulleistung besitzen können (Kingston & Nash 2011; Black & Wiliam 1998a). Es existieren jedoch nach wie vor Lernbereiche, zu welchen nur eingeschränkt belastbare Erkenntnisse vorliegen (Dunn & Mulvenon 2009; Klauer 2014). Die Studie knüpft an diese aktuelle Befundlage an und untersucht die Effekte eines testbasierten Förderkonzepts mit integrierten Selbsteinschätzungen (z.B. Andrade & Valtcheva 2009; Saddler & Good 2006) für Schreibaufgaben im Deutschunterricht (z.B. Underwood & Tregidgo 2006). Konkret sollten vor allem zu folgenden drei Primärforschungsfragen Erkenntnisse gewonnen werden:

- F1: Besitzen formative Leistungsmessungen bei ganzheitlichen Leistungen einen positiven Effekt auf die Kompetenzausprägung von Schülerinnen und Schülern bei Schreibaufgaben?
- F2: Besitzen Selbsteinschätzungen im Rahmen formativer Leistungsmessungen bei ganzheitlichen Leistungen einen positiven Effekt auf die Kompetenzausprägung von Schülerinnen und Schülern bei Schreibaufgaben?
- F3: Nähern sich die Selbsteinschätzungen der Schülerinnen und Schüler zur ganzheitlichen Leistung „Inhalte zusammenfassen“ durch Übung und Feedback dem Urteil des Lehrenden an?

Die Hypothesenprüfungen erfolgten im Rahmen einer quasiexperimentellen Feldstudie, welche an einer bayerischen Realschule in der neunten und zehnten Klasse ($n = 235$) zur Kernkompetenz „Inhalte zusammenfassen“ im Deutschunterricht durchgeführt wurde. Das Experiment sah vor, dass alle Schülerinnen und Schüler zunächst eine Kurzzusammenfassung zu einem journalistischen Text schreiben. Anschließend wurde das Sample in Experimental- und Kontrollgruppe geteilt. Die Experimentalgruppe führte nun anhand eines Erwartungshorizontes sowie eines Bewertungsbogens eine Selbsteinschätzung zur angefertigten Inhalts-zusammenfassung durch. Die Kontrollgruppe erledigte in dieser Zeit eine alternative Aufgabe mit Text- und Kompetenzbezug. Der Bewertungsbogen bestand aus elf Kriterien, welche die Unterrichtsinhalte zu dieser Aufsatzgattung proportional abbilden und welche ausschließlich bestehenden Bewertungsinstrumenten oder zugelassenen Lehrwerken entnommen wurden. Die Lehrkräfte korrigierten nach der Arbeitsphase von beiden Gruppen die Ausarbeitungen und verbesserten im Zuge dessen auch die Selbsteinschätzungen der Testgruppe. Alle Schülerinnen und Schüler erhielten durch die Herausgabe sowie die Besprechung im Unterricht eine Rückmeldung zu ihrer schriftlichen Ausarbeitung, die Testgruppe durch die Rückgabe des zuvor ausgefüllten Bewertungsbogens zusätzlich Hinweise auf die getätigte Selbsteinschätzung.

Die Auswertung ergab (F1), dass sich die teilnehmenden Schülerinnen und Schüler (Experimental- und Kontrollgruppe) von Pre- zu Posttest signifikant ($p = .000$) steigern konnten: Das arithmetische Mittel der Bewertungseinheiten (max. 44 BE) stieg von 24.9 ($s = 5.46$) auf 33.2 ($s = 4.30$) an. Damit konnte der Nachweis erbracht werden, dass formative Leistungsmessungen auch bei Schreibaufgaben im Deutschunterricht große Effekte ($d_c = 1.68$) auf die Kompetenzausprägung besitzen können. Der Vergleich der Leistungsentwicklung von Experimental- und Kontrollgruppe (F2) zeigte, dass das Durchführen von Selbsteinschätzungen in der vorliegenden Studie keine bedeutsame Wirkung ($F = .615$; $p = .434$) auf die Ausprägung der Kompetenz eine Inhaltzusammenfassung erstellen zu können besaß. Unterhalb des Signifikanzniveaus konnte über einen Vergleich der Effektstärken jedoch festgestellt werden, dass sich die

Experimentalgruppe ($d_e = 1.77$) stärker verbesserte als die Kontrollgruppe ($d_e = 1.50$). Durch die Ergebnisauswertung wird weiter ersichtlich (F3), dass das mehrmalige Üben von Selbsteinschätzungen dazu führt, dass sich das Urteil der Lernenden der Bewertung der Lehrkräfte annähert. Dies belegt in erster Linie die hoch signifikante ($p = .000$) Reduktion der Abweichung von durchschnittlich 9.5 BE ($s = 5.51$) beim Pretest auf durchschnittlich 4.4 BE ($s = 3.43$) zum Zeitpunkt des Posttests. Die geringen Unterschiede zwischen Lehrer- und Schülerurteil bei der finalen Messung zeigen, dass die Lernenden die Kompetenz des (Selbst)Einschätzens erlernen können, was mit Blick auf selbständige Lernprozesse eine essentielle Kompetenz darstellt.

1 Einleitung

Im aktuellen Diskurs der Schullandschaft nehmen Schlagwörter wie „*Differenzierung*“ (Bönsch 2012; Sorrentino et al. 2010; Müller 2012; Haag & Streber 2014; Graumann 2002), „*Inklusion*“ (Hinz 2002; Pfahl 2012) oder „*Umgang mit Heterogenität*“ (Grunder & Gut 2012; Gommola 2009; Trautmann & Wischer 2012; Sturm 2013; Seitz 2007) eine bedeutsame Stellung ein. Hinzu kommen Forderungen nach „*prozessorientierten Leistungsermittlungen*“ sowie einer „*stärkeren Berücksichtigung individueller Lernentwicklungen*“ (Vollstädt 2005; Brenner 2006). Auch die Kritik an den bestehenden, meist auf Selektion ausgerichteten, Leistungsmessungen verklingt nach wie vor nicht (Schusser 2004; Oelkers 2002).

Die Auseinandersetzung mit diesen zentralen Themenfeldern der Erziehungswissenschaft hat zur Folge, dass im Bereich der schulischen (Leistungs)Diagnostik eine Neuorientierung Einzug hält: Wer mit Blick auf eine immer heterogener werdende Schülerschaft individuelle Erkenntnisse gewinnen und Differenzierungsmaßnahmen einleiten möchte, benötigt Diagnoseverfahren, welche jene notwendigen Informationen auch liefern können (Paradies et al. 2009; Winter 2002). Herkömmliche traditionelle Leistungserhebungen (allein) sind hier oftmals überfordert. Um präzise und vielschichtige Messungen durchführen zu können, sind innovative, Leistungsunterschiede besser verdeutlichende diagnostische Maßnahmen und Konzepte von Nöten. Auch die Rückmeldeoptionen bzw. ganz allgemein Feedbackmaßnahmen haben im Zuge veränderter Vorstellungen von schulischen Leistungserfassungen einen Wandel durchlaufen (Bastian et al. 2007; Lütger 2000). Innerhalb der Diskussion, wie zeitgemäßer Unterricht im 21. Jahrhundert aussehen sollte, hat die deutlichere *Orientierung an Kompetenzen* eine Schlüsselstellung eingenommen (Weinert 2002; Altrichter & Schratz 2004). Diese Relevanzverschiebung wird auf schulpraktischer Ebene vor allem an der Festsetzung und Implementation bundesweit einheitlicher Bildungsstandards sowie an der Konzeption neuer, kompetenzorientierter Lehrpläne ersichtlich (Klieme et al. 2009). Im Wissenschaftsbetrieb an Hochschulen und Universitäten zeigt sich die größere Bedeutung von Kompetenzen unter anderem an der Erstellung von (neuen) Kompetenzmodellen (z.B. Leutner et al. 2013; Wilhelm & Nickolaus 2013; Jordan et al. 2012) oder an der Auseinandersetzung mit der auch forschungsmethodisch bedeutsamen Frage, wie Kompetenzen überhaupt diagnostisch zu fassen sind (z.B. Schwabe et al. 2012; Weinert 2002; Bögelholz & Eggert 2013). Zu den primären Zielen der Neuausrichtung zählt vor allem die Fähigkeit, „selbstständig zu arbeiten“ (Baumert et al. 2003) bzw. im Kontext der Schule „selbstständig zu lernen“ (Traub 2011; Killus 2009; Konrad & Traub 2009; Boekaerts 1999; Schiefele & Pekrun 1996). All diese Grundintention umsetzende Ansätze eint, dass der Lernende individuell gefördert werden soll und als Subjekt wieder im Mittelpunkt der Betrachtung steht (Held 2000; Ludwig 1999). Begriffe wie „subjektorientiertes Lernen“ (Hempel 1999; Holzbrecher 1999b) oder „subjektorientierte Didaktik“ (Ludwig 2005; Holzbrecher 1999a; Heursen 1998) verdeutlichen diese Rückbesinnung auf einen subjektorientierten Unterricht (Ramseger 1991).

Deutlich vielschichtiger sind die Reaktionen auf die allgegenwärtige Kritik an der Form oder gar der Existenz von Zensuren als Selektionsinstrument an deutschen Schulen (z.B. Sacher 2002). In diesem Zusammenhang muss stets auch auf die tiefer ansetzende Debatte um einen (veränderten) Leistungsbegriff verwiesen werden (v. Saldern 2011; Winter 2010; Jürgens 2000).

Als Konsequenz der angerissenen Neuorientierungen im deutschen Schulwesen, insbesondere im weitläufigen Feld der Leistungsmessungen, kann in erster Linie die Entwicklung innovativer Konzepte zur (Leistungs)Diagnostik angeführt werden (z.B. Gläser-Zikuda & Hascher 2007; Bohl

2005; Höhmann 2002). Wie bereits skizziert, ist auch für immer stärker geforderte Differenzierungsmaßnahmen ein Umdenken bei der schulischen (Leistungs-)Messung notwendig. Die folglich verlangte Neuausrichtung der Diagnostik im Rahmen von schulischen Lehr-Lernprozessen spiegelt sich sowohl in zahlreichen wissenschaftlichen als auch eher der Praxisliteratur zu zuordnenden Publikationen wider. Schlagwörter dieser diagnostischen Neuausrichtung sind Termini wie „Neue Prüfkultur“ (Sacher 2009a; Stern 2008; Stern 2004), „Formative Leistungsmessung“ (Black & Wiliam 1998a; McDonald & Boud 2003; Noonan & Duncan 2005; Maier 2010), „Alternative Leistungsmessungen“ (z.B. Winter 2010; Lütgert 1999; Lissmann 2010; Grunder & Bohl 2004), „Lernverlaufsdiagnostiken“ (Klauer 2014; Mand 2003; Wilbert & Linnemann 2001) oder schlicht „Neue Formen der Leistungsbewertung“ (z.B. Bohl 2004; Grittner 2010). Auch der Begriff bzw. das Konzept der „Pädagogischen Diagnostik“ (Ingenkamp 1999; Ingenkamp & Lissmann 2008) erlebt im Zuge dieser Neuformierung eine Renaissance.

Die vorliegende Arbeit knüpft an die genannten Komponenten des aktuellen pädagogischen Diskurses an und verbindet die Gedanken zur Entwicklung bzw. Messung von Kompetenzen sowie zu innovativen Konzepten der schulischen Leistungsdiagnose. Konkret ist die Studie in eine Umsetzung des Konzeptes der formativen Leistungsmessung eingebettet und untersucht zum einen die Fähigkeit der Schülerinnen und Schüler sich „selbst einschätzen zu können“ und zum anderen die Kompetenzentwicklung der fächerübergreifend bedeutsame Fähigkeit „Inhalte zusammenfassen und wiedergeben können.“

Die einen zentralen Part im Experiment einnehmenden Selbsteinschätzungen spielen in vielen zeitgemäßen Überlegungen zur Leistungsmessung an Schulen eine bedeutsame Rolle (z.B. Winter 2010; Paradies et al. 2009), nicht zuletzt, da im Zuge der wachsenden Bedeutung selbständiges Lernen zu fördern bzw. gar auszubilden, die Kompetenz Lernstände eigenständig zu erfassen als unverzichtbar deklariert werden muss (z.B. Bräu 2002; Beck et al. 1995; Bannach 2002). Hinzu kommt, dass Schreibleistungen, welche vor allem im Aufsatzunterricht nicht nur gefordert sondern auch bewertet werden, im Feld der schulischen Leistungsmessung ohnehin eine besondere Rolle einnehmen (z.B. Jürgens & Sacher 2008; Altenburg 2002; Sacher 2009a); immer wieder wird in diesem Zusammenhang diskutiert, inwieweit sogar geschulte Lehrkräfte in diesem Spezialfall der Schulleistungsbewertungen die notwendige diagnostische Kompetenz besitzen, Aufsätze korrekt und vergleichbar einzuschätzen (z.B. Karing et al. 2011). Aus diesem Grunde ist zu prüfen, inwieweit Schülerinnen und Schüler diesen diagnostischen Vorgang selbstständig durchführen können (z.B. Winter 1991; Viebahn 1982; Winter 2000b; Filene 1969; Buschmann 2001) und ob sie in der Lage sind, ihre Kompetenz, sich selbst einzuschätzen, steigern können.

Primärziel der Studie ist es, innerhalb des abgesteckten Rahmens durch eine Feldstudie weitere aktuelle empirisch gestützte Erkenntnisse zu liefern. Konkret stellen die Ergebnisse einen Beitrag zur Erforschung der Effekte formativer Leistungsmessungen als Rückmeldemöglichkeit während bestehender Lehr-Lernprozesse auf die Schulleistung bei Schülerinnen und Schülern dar. Die Untersuchung wird im Fach Deutsch im Rahmen des schriftlichen Sprachgebrauchs (Aufsatzform Inhaltsangabe) in der Sekundarstufe I einer weiterführenden Schule durchgeführt. Bestandteil der formativen Leistungsmessungen sind neben den schriftlichen Inhaltszusammenfassungen auch Selbsteinschätzungen zu den angefertigten Kurzaufsätzen. Auf diesem Wege sollen Informationen darüber gewonnen werden, welche Effekte diese auf die Schulleistung besitzen. Zudem kann allgemein eine Aussage darüber getroffen werden, ob sich Schülerinnen und Schüler auch bei so genannten „ganzheitlichen Leistungen“ (Sacher 2009a) ähnlich wie die Lehrkräfte einschätzen (Winter 1991). Im Zuge dieser Analyse ist es weiter möglich, prinzipiell zu eruieren, ob von den Lernenden dieser Altersstufe auch im Aufsatzunterricht realistische Selbsteinschätzungen vorgenommen werden können.

2 Theoretische Grundlagen

2.1 Formative Leistungsmessungen

2.1.1 Einführung in die Thematik

Im Zuge der Debatte um eine „Neue Lernkultur“ (Sacher 2009a; Winter 2010) wurden auch schulische Leistungsmessungen neu diskutiert (z.B. Herold & Landherr 2003; Müller 2010). Schon der Blick auf eines der Primärziele der „Neuen Lernkultur“, die Berücksichtigung der *individuellen Lernvoraussetzungen* und *-prozesse* der Menschen (Müller 2010), zeigt, dass zeitgemäße Lernstandsüberprüfungen in der Lage sein müssen, die Individualität der Schülerinnen und Schüler auch zu fassen (Kunze 2009). Somit fordert eine „Neue Lernkultur“ mit ihren veränderten Methoden und Unterrichtsformen auch ein Umdenken hinsichtlich der Bewertung von Leistung und der Konzeption von Prüfungen (z.B. Winter 2010; Sacher 2009a; Jürgens & Sacher 2008; Huber 2002; Becker 2005). Unter dem Schlagwort „Neue Prüfkultur“ sammeln sich eine Vielzahl an Konzepten und innovativen Ideen für die verlangte Neuausrichtung (z.B. Lissmann 2010; Schusser 2004; Oelkers 2002). Innerhalb dieser Strömung spielen drei Komponenten, welche in anderen Ländern schon lange Bestandteile von Qualitätsoffensiven im Schulsystem sind (z. B. Stiggins 2005), eine bedeutsame Rolle (Vollstädt 2005):

- Leistungsermittlung und -bewertung beziehen sich nicht nur auf Ergebnisse und Produkte der Lerntätigkeit der Schülerinnen und Schüler, sondern schließen vermehrt auch *prozessuale Komponenten* der Schülertätigkeit ein.
- Rückmeldungen des Lernfortschritts müssen auch *in bewertungsfreien Lernphasen* in den Unterricht integriert sein. In der Regel werden diese nicht wie selektive Leistungsmessungen erst ans vorläufige Ende eines längeren Lernprozesses gesetzt, sondern begleiten die Schülerinnen und Schüler während des gesamten Lehr-Lernprozesses.
- Bei derartigen Rückmeldungen während des Lernprozesses geht es nicht nur um den (kriterialen) Vergleich mit curricularen Maßstäben, sondern auch um *individuelle Lernentwicklungen* sowie eine Analyse aufgetretener Probleme.

Hinzu kommt, dass Begriffe wie „*Differenzierung*“ (Bönsch 2012, Sorrentino et al. 2010), „*Inklusion*“ (Hinz 2002; Pfahl 2012) oder „*Umgang mit Heterogenität*“ (Gomolla 2009; Trautmann & Wischer 2012) den aktuellen Diskurs in der (deutschen) Bildungslandschaft entscheidend mitbestimmen. Sollen individuelle Lernstände zur Optimierung weiterer Lehr-Lernprozesse diagnostiziert werden, sind neue, verbesserte Messinstrumente und Methoden notwendig (Prenzel et al. 2009; Stiggins 2005; Yorke 2003; Dochy 2001). Winter (2010, S. 29) sieht prinzipiell ein eklatantes „Spannungsverhältnis“ zwischen Grundmerkmalen einer „neuen Lernkultur“ (z.B. Selbständigkeitsentwicklung, Prozessorientierung, Demokratisierung oder situiertes Lernen) und einer „Leistungsmessung nach traditionellen Gesichtspunkten“.

Die auszugsweise angeführten Aspekte finden ihre Berücksichtigung jedoch auch in zahlreichen anderen innovativen Konzepten zur Leistungsmessung, die beispielsweise als „alternative Leistungsmessungen“, „Lernverlaufsdiagnostiken“, „Pädagogische Diagnostik“ oder schlicht „neue Formen der Leistungsmessung“ deklariert werden.

In allen angeführten Konzepten werden Messungen und Lernstandserhebungen in erster Linie im Sinne des Lernfortschritts vorgenommen (Swearingen 2002). Teil dieser Auflistung an Termini, die allesamt für eine Entwicklung der Diagnosepraktiken stehen, muss auch der Begriff

„*formative Leistungsmessung*“ („*assessment for learning*“) sein, der im deutschsprachigen Raum jedoch nach wie vor eher selten eingesetzt wird und daher vor allem bei Schulpraktikern wenig bekannt ist (Klauer 2014).

Anders verhält es sich bei dem seit jeher verwendeten Terminus „*Förderdiagnostik*“, der sowohl in eher wissenschaftlichen (z.B. Breitenbach 2003; Mutzeck & Jogschies 2004; Bundschuh 2007) als auch eher praxisnahen Publikationen Verwendung findet (z.B. Wendeler 2000; Oltmanns 2006). Problematisch ist in diesem Kontext, dass der Begriff partiell synonym zu „*formativen Leistungsmessungen*“ verwendet wird, teilweise jedoch auch einen besonderen Bereich der Diagnostik zur Verbesserung des Lernens abdecken soll (Klauer 2014; Rittmeyer 2009; Becker 2005). Exemplarisch kann der breite Einsatzbereich des Terminus daran gezeigt werden, dass Förderdiagnostik zum einen in sämtlichen Sparten der *Sonder- bzw. Förderpädagogik* (z.B. Bundschuh 2007; Eberwein & Knauer 2003), zum anderen für die Ermittlung von Stärken und Schwächen bei Schülerinnen und Schülern an *Regelschulen* (z.B. Wendeler 2000; Dinges 2002) Anwendung findet. Dies hat zur Folge, dass nicht nur eine einheitliche Definition von Förderdiagnostik schwer möglich ist (Popham 2006), sondern auch nicht eindeutig festgelegt werden kann, ob „*formative Leistungsmessung*“ und „*Förderdiagnostik*“ das gleiche Spektrum abdecken oder unterschiedliche diagnostische Vorgänge und Konsequenzen bezeichnen.

Weiterhin besteht das Problem, dass vor allem in der Schulpraxis eine strikte Trennung zwischen den bereits angeführten Begriffen „*alternative Leistungsmessung*“, „*formative Leistungsmessung*“ oder „*Pädagogische Diagnostik*“ kaum oder nur unzureichend stattfindet. Häufig werden alle Konzepte oder Formen einer „Neuen Prüfkultur“ gleichgesetzt, was zu irrtümlichen Vorstellungen führt und viele Konzepte aus diesem Grunde keinen Einzug in den Unterrichtsalltag finden. Hauptanliegen dieses Kapitels ist es daher, eine übergeordnete Verortung des für diese Arbeit zentralen Begriffs „*formative Leistungsmessung*“ sowie eine Abgrenzung zu anderen bedeutenden Termini der schulischen (Leistungs)Diagnostik vorzunehmen. Anschließend soll ein Einblick in mögliche Konzepte zur Umsetzung „*formativer Leistungsmessungen*“ ermöglicht werden und der aktuelle Forschungsstand kurz skizziert werden.

2.1.2 Formative Leistungsmessung: Annäherung, Definition, Erläuterung und Abgrenzung

2.1.2.1 Definition anhand diagnostischer Strategien

Wie bereits erwähnt, eint alle Ansätze zu „*formativer Leistungsmessung*“, dass sie auf die „*Verbesserung von Lernen*“ und nicht auf die Beurteilung von Endprodukten zum Zwecke der Selektion“ abzielen (McDonald & Boud 2003, S. 2010).

Diese simple Grundüberlegung, welche als übergreifende Gemeinsamkeit bezeichnet werden kann, zeigt jedoch bereits, dass eine konkretere Verortung von „*formativer Leistungsmessung*“ mehrere Definitionskriterien notwendig werden lässt: Schon die Bestimmung von McDonald & Boud (2003) basiert auf *zwei* verortenden Aspekten: Einer Aussage zur *Grundintention* der Messung (*Selektion* oder *Modifikation*) und einer zum *Gegenstand* des Diagnosevorgangs (*Endprodukt* oder *Lernschritt bzw. Lernprozess*).

Ähnlich agiert auch Maier (2013, S. 260), der „*formative Leistungsmessung*“ folgendermaßen verortet:

„Formative Leistungsmessung bedeutet, dass man über die gesamte Lehr-Lernsequenz hinweg immer wieder prüft, was Schüler(innen) tatsächlich gelernt haben. Wenn die Erfassung von Lernergebnissen erst den Abschluss des Lehr-Lernprozesses darstellt (summativ Produktevaluation für die Notengebung), kann nicht mehr reagiert werden.“

Die Erläuterung von Maier (2013) beinhaltet ebenfalls zwei Kategorisierungskriterien, sie verknüpft den Zweck (*formativ* oder *summativ*) mit dem Zeitpunkt einer Leistungsmessung (*während des Lernprozesses* oder *am Ende des Lehr-Lernprozesses*). Andere aktuelle Definitionsvarianten nehmen diese Koppelung von Intention und Zeitpunkt nicht vor.

Becker (2005, S. 9f.) beispielsweise verwendet die Begriffe „formative“ bzw. „summative Evaluation“ als Synonym für „Prozessevaluation“ bzw. „Produktevaluation“:

„Bei der formativen Evaluation *oder* Prozessevaluation werden die einzelnen Teilergebnisse im Beurteilungszeitraum registriert und verknüpft, z.B. in aufsteigender Tendenz die Noten 5,4,3,2,1 oder in absteigender 1,2,4,6 [...]. Die summative Evaluation oder die Produktevaluation bezieht sich auf die am Ende des Beurteilungszeitraumes registrierbare Gesamtleistung, z.B. auf die Note drei im Fach Englisch...“

Die obige Definition von Becker (2005) berücksichtigt die intentionale Ebene nicht, alle seine Beispiele sind aus dem Bereich der Selektionsdiagnostik gewählt. Um zwischen „Feedback-“ und „Selektionsfunktion“ zu differenzieren, verwendet Becker (2005, S. 23) (zusätzlich) die Begriffe „Ergebniskontrolle“ und „Leistungsbeurteilung“.

Schon die drei angeführten Verortungen des Terminus „formativ“ (meist in Abgrenzung zum Begriff „summativ“) haben gezeigt, dass dieser durchaus unterschiedlich verstanden wird (McDonald & Boud 2003; Maier 2013; Becker 2005). Die Heterogenität der Definitionsvarianten nimmt mit einer Erhöhung der bestimmenden Aspekte bzw. Komponenten gar noch zu (Klauer 2014), was vor allem bei komplexeren und spezifischeren Festlegungen ersichtlich wird (z.B. Black & Wiliam 1998a; Sadler 1989; Boston 2002).

Jene einleitend dargelegten Gedanken zu einer Definition sowie die Suche nach prinzipiellen Gemeinsamkeiten und Charakteristika von „formativen Leistungsmessungen“ führen zur Grundverortung auf eine *theoretisch-normative Ebene* der Diagnostik, die über Kriterien und diagnostische Entscheidungsfelder Messvorgänge bestimmt (z.B. Eid & Petermann 2006; Leutner 2001; Amelang & Zielinski 2004; Schmidt-Atzert & Amelang 2012; Kubinger 2009).

Einteilungen erfolgen hier in der Regel über so genannte „*diagnostische Strategien*“ oder „*diagnostische Dimensionen*“ (Pawlik 1976; Amelang & Zielinski 2004; Eid & Petermann 2006). Jäger (1986, S. 67) definiert „diagnostische Strategien“ [oder „diagnostische Dimensionen“] als „auf diagnostische Daten aufbauende Konzeptionen, mit deren Hilfe der Diagnostiker sein antizipiertes Ziel zu erreichen sucht.“ So wird in der Evaluationsforschung spätestens seit Scriven (1967) zwischen „summative and formative program evaluation“ unterschieden.

Eine „Evaluation“, also ganz allgemein das „Bewerten von Sachverhalten“ (Wottawa 2001, S. 152), kann somit aus *zwei* gegenüberstehenden Perspektiven bzw. mit zwei unterschiedlichen Zielsetzungen vorgenommen werden: *Treatment-vorbereitend* und/oder *Treatment-abschließend* versus *Treatment-begleitend*.

Bloom et al. (1971) erweitern diese zunächst nur auf Evaluationsverfahren bezogene Unterscheidung auf prinzipiell zwei verschiedene Formen der Beurteilung: Eine „summative Beurteilung“ findet *nach* einem Lernprozess statt, um zu überprüfen bzw. zu bestimmen ob dieser stattfand; eine „formative Form der Beurteilung“ wird zur Förderung bzw. Unterstützung – nicht zur Selektion oder Klassifizierung – des Lernenden *während* eines Lernvorgangs platziert. Pawlik (1976) strukturiert diese ersten definierenden Ansätze dahingehend, dass er über „Modell und Praxisdimensionen“ die psychologische Diagnostik nach Zielsetzungen möglichst differenziert unterteilt. Während Scriven (1967) und Bloom et al. (1971) den Dimensionen „formativ“ bzw. „summativ“ gleichzeitig eine Intention *und* einen Zeitpunkt, zu welchem die

Messung stattfindet, zuordnet, trennt Pawlik (1976) diese Kopplung wieder, da für ihn in einem *ersten Schritt* eine „angewandt-psychologische Interventionsstrategie“ gewählt wird und anschließend in einem *zweiten Schritt* anhand „vier Alternativdimensionen psychodiagnostischer Zielsetzung“ weitere Zuordnungen möglich sind. Nach Pawlik (1976, S. 16) bestehen hinsichtlich der Zusammenstellung der Kriterien „verschiedenste [...] Kombinationen“, wobei „Änderungen in der Strategie angewandt-psychologischer Interventionen auch Konsequenzen für die Entwicklung der psychologischen Diagnostik haben können“ (S. 18). Vereinfacht können die Einteilungsstrategien und Dimensionen nach Pawlik (1976) folgendermaßen dargestellt werden (Abb. 1):

I	angewandt-psychologische Interventionsstrategie		Selektionsstrategie	↔	Modifikationsstrategie
II	Alternativdimensionen psychodiagnostischer Zielsetzung	1	Statusdiagnostik	↔	Prozessdiagnostik
		2	normorientierte Diagnostik	↔	kriterienorientierte Diagnostik
		3	Testen	↔	Inventarisieren
		4	Diagnostik als Messung	↔	Diagnostik als Information

Abb. 1: Modell- und Praxisdimensionen psychologischer Diagnostik nach Pawlik (1976)

Ausgangspunkt für alle weiteren Bestimmungen ist folglich die erste Einteilung in „*Selektions-* versus *Modifikationsstrategien*“, wobei als „Selektionsstrategie eine Vorgehensweise bezeichnet wird, die eine Optimierung durch geeignete Auswahl von Personen und/oder Bedingungen sucht“ (Pawlik 1976, S. 15). Im Bereich der „pädagogischen Selektion“ wird von „vorgegebenen Bedingungen ausgegangen“, das bedeutet, üblicherweise findet eine Personenselektion statt (Pawlik 1976, S. 15). Demgegenüber „versteht man als Modifikationsstrategie eine Vorgehensweise, die eine Veränderung des Verhaltens und/oder von Bedingungen sucht“ (Pawlik 1976, S. 16). Im Kontext von schulischen Lehr-Lernprozessen liegen in der Regel Verhaltensmodifikationen vor (Pawlik 1976; Leutner 2001). Mit Blick auf die Praxis stellt Pawlik (1976) jedoch fest, dass sich viele der möglichen Strategien selten als Rein- sondern eher als *Mischformen* verstehen, wobei die Pole jeweils die Endpunkte eines Strategiekontinuums darstellen. Diese Feststellung ist von besonderer Bedeutung für die Interventionsentscheidung, da es auf der operationalen Ebene zwischen „reinen Selektionsstrategien und reinen Modifikationsstrategien die verschiedensten Abstufungen von Mischstrategien gibt“ (Pawlik 1976, S. 16). Dies betrifft auch im Zuge einer weiteren Bestimmung die Ebene der „Alternativdimensionen psychologischer Zielsetzungen“. Betrachtet man nicht das im Vergleich zur pädagogischen Psychologie (noch) breitere Spektrum der klinischen Psychologie, sondern verweilt im pädagogischen Anwendungsbereich, so ist vor allem die erste Zielsetzung entscheidend: Hier wird zwischen *Status-* und *Prozessdiagnostik* unterschieden (auch Eid & Petermann 2006; Leutner 2001). „Statusdiagnostik bezeichnet dabei eine psychologische Diagnostik mit dem Ziel, einen psychologischen Ist-Zustand festzustellen“, sie „geht davon aus, dass das diagnostische Ergebnis, der Befund, [...] hohe Stabilität besitzt“ (Pawlik 1976, S. 24). „Prozessdiagnostik“ auf der anderen Seite besitzt das Ziel, „Veränderungen in psychologischen Variablen festzustellen“ (Pawlik 1976, S. 24). Wie bereits angeführt, sieht Pawliks (1976) „Diagnose der Diagnostik“ eine *Modell-* oder *Theoriedimension* und eine *Praxisdimension* vor. Auf der Ebene der theoretischen Verortung sind die „angewandt-psychologischen Interventionsstrategien“ und die vier „Alternativdimensionen

psychodiagnostischer Zielsetzung“ unabhängig voneinander (Scriven 1967, 1996). Vor allem mit Blick auf die Praxis „hängt die erste Zielalternative [*Status-* versus *Prozessdiagnostik*] [jedoch] eng mit der Unterscheidung zwischen Selektions- und Modifikationsstrategien zusammen“ (Pawlik 1976, S. 24), was sich an den bereits einführend aufgenommenen Definitionen von Bloom et al. (1971), McDonald & Boud (2003) und Maier (2013) zeigt. Bringt man nach Scriven (1967) die Begriffe „formativ“ und „summativ“ in diesen Kontext, dann kann folglich hinter diesen *ausschließlich* eine Intentionsangabe (*Selektions-* versus *Modifikationsstrategie*) oder eben eine Intentionsangabe in Verbindung *mit* einer Messung von Veränderungen (*Prozessdiagnostik*) oder Ist-Zuständen (*Statusdiagnostik*) verstanden werden – beide Auffassungen sind in Theorie und Praxis anzufinden. Leutner (2001) weist zurecht daraufhin, dass vor allem im Bereich der „Evaluationsforschung“ (z.B. Henry et al. 2013; Wholey 1996) sogar eine weitere Belegung von „formativ“ und „summativ“ möglich ist: So liefern die Termini „formativ“ und „summativ“ in vielen Ausführungen zur Evaluationstheorie keine Aussage über den Verwendungszweck eines Messergebnisses (*Selektion* oder *Modifikation*), wohl aber über den Zustand des zu messenden Gegenstandes (*stabil* versus *veränderbar*) und den Zeitpunkt der Messung (*während eines Treatments* versus *nach einem Treatment*) (Scriven 1967; Wholey 1996). Leutner (2001, S. 523) fasst dies konkret so zusammen, dass „eine summative Evaluation [dann] vorliegt, wenn Statusdiagnostik am Ende einer Lehr-Lerneinheit stattfindet, wogegen Prozessdiagnostik einer formativen Evaluation entspricht.“

Zusammenfassend kann bis zu diesem Stand somit bilanziert werden, dass die Formulierung „formative Messungen“ *Ziel, Gegenstand* und *Zeitpunkt* einer Erhebung, partiell aber auch nur eine dieser Dimensionen festlegen kann. Versucht man die potenziellen Verbindungen zwischen den Strategien weiter zu ergünden, kann auf der theoretischen Ebene aller diagnostischen Vorgänge von Pawlik (1976) der Zusammenhang zwischen Interventionsstrategien (*Selektions-* versus *Modifikationsstrategie*) und primärer Alternativdimension (*Selektions-* versus *Prozessdimension*) dahingehend präzisiert werden, dass „Selektionsstrategien in der Regel mit einer reinen Statusdiagnostik zurechtkommen, Modifikationsstrategien [...] sowohl Verfahren der Prozessdiagnostik als auch statusdiagnostische Methoden [...] voraussetzen“ (S. 24).

2.1.2.2 Definition über die Abgrenzung fachverwandter Termini

„Formative Leistungsmessung“ versus „Förderdiagnostik“: Mit Blick auf den skizzierten Zusammenhang erscheint es sinnvoll, auf das partiell bestehende und bereits verwiesene Problem der inhaltlichen Belegung des Begriffes „Förderdiagnostik“ kurz einzugehen. Durch die angeführten diagnostischen Dimensionen wird ersichtlich, dass der Terminus „Förderdiagnostik“ auf der Ebene der theoretischen Bestimmung diagnostischer Prozesse entweder als Synonym für „Modifikationsdiagnostik“ (Pawlik 1976; Scriven 1967) oder zur Erweiterung des diagnostischen Spektrums eingesetzt wird (Bloom et al. 1971). Verlässt man die Ebene der diagnostischen Strategien können einige gravierendere Unterschiede zwischen den Begriffen „formative Leistungsmessung“ und „Förderdiagnostik“ ausgemacht werden (Klauer 2014).

Ausgangspunkt dieser ist die unterschiedliche Herkunft der Termini. Der vor allem im deutschsprachigen Bereich verwendete Begriff „Förderdiagnostik“ deckt ein sehr breites Einsatzspektrum ab, weshalb eine Definition, die allen förderdiagnostischen Konzepten eine theoretische Grundlage bietet, kaum möglich ist (z.B. Breitenbach 2003). Exemplarisch kann das breite Einsatzspektrum daran gezeigt werden, dass der Terminus „Förderdiagnostik“ zum einen diagnostische Prozesse im Bereich der Sonderpädagogik bezeichnet (z.B. Eberwein & Knauer 2003; Bundschuh 2007, 1998), zum anderen aber auch alle Maßnahmen zur gezielten Erfassung von

Stärken und Schwächen bei Lernenden an Regelschulen in sich subsumiert (z.B. Schmischke 2008; Mand 2003; Wendeler 2000; Dinges 2002). Noch weiter wird die Belegung der Formulierung „Förderdiagnostik“, wenn es nicht mehr nur um Grundintentionen und Anwendungsfelder, sondern um Umsetzungsoptionen geht (Popham 2006; Klauer 2014). Anders verhält es sich bei dem internationalen bzw. aus den Vereinigten Staaten stammenden Begriff „formative assessment“. Dieser Terminus ist trotz aller auch hier vorhandenen Heterogenität enger besetzt und zumindest auf theoretischer Ebene konkreter (Klauer 2014). So liegt allen Konzepten von „formativer Leistungsmessung“ das Ziel einer „mehrperspektivischen Erfassung [...] und Optimierung von Schülerleistungen“ zugrunde (Maier 2010, S. 294).

„Formative Leistungsmessung“ versus „Alternative Leistungsmessung“: In der Schulpraxis findet sich häufig die Vorstellung, „formative Leistungsmessungen“ und „Alternative Leistungsmessungen“ bezeichneten den gleichen Gegenstand oder seien zumindest untrennbar miteinander verknüpft. So wird in der Regel meist davon ausgegangen, dass bei der Umsetzung einer formativ genutzten Leistungsmessung auch auf innovative (alternative) Formen der Leistungsmessung zurückgegriffen wird. Diese Sichtweise kann aufgrund des gemeinsamen Hintergrunds einer „neuen Prüfkultur“ durchaus nachvollzogen werden, muss mit Blick auf die bereits vorgenommene theoretische Verortung des Begriffs „formative Leistungsmessung“ jedoch in dieser Eindeutigkeit abgelehnt werden (Pawlik 1976; Scriven 1967; Leutner 2001; Eid & Petermann 2006). Unter Berücksichtigung der theoretischen Modelldimension (Pawlik 1976) ist es unstrittig, dass die in der Schulpraxis durchaus vorkommende Gleichsetzung der Termini „formative Leistungsmessung“ und „alternative Leistungsmessung“ nicht zulässig ist: Während der Begriff „formativ“ die Primärintention bzw. die primäre *Intentionsstrategie* einer Messung vorgibt (Pawlik 1976), legt die Formulierung „alternativ“ die *Form* der Lernstandserhebung fest (Sacher 2009a). Weiter müssen bei „summativen“ Leistungsmessungen meist eine Reihe von schulrechtlichen Vorgaben berücksichtigt werden (Paradies et al. 2009; Woolfolk 2012), was die Lehrkräfte nur langsam von traditionellen Klassenarbeiten abrücken lässt. Nur allmählich werden auch alternative Formen der Leistungsmessung für selektierende Diagnostikvorgänge genutzt (Bohl 2004b; Lissmann 2007). Diese im Unterrichtsaltag häufig (noch) bestehende Situation hat zur Folge, dass „formative Leistungsmessungen“ mit „alternativen Formen“ und „summativen Leistungsmessungen“ mit eher „traditionellen Formen der Leistungsmessung“ in Verbindung gebracht werden (Bohl 2004b; Winter 2007). Wie beschrieben ist dies so jedoch nicht zwingend der Fall: Auf der Ebene von diagnostischen Strategien und Dimensionen legt die *Absicht* einer Leistungsmessung nicht automatisch die *Form* dieser fest (z.B. Scriven 1996; Woolfolk 2012). Dies zieht für eine Umsetzung nach sich, dass „formative Leistungsmessungen“ unterschiedliche Formen einnehmen können (Grunder & Bohl 2004; Lissmann 2010; Paradies et al. 2009; Maier 2014b), auf der Ebene der theoretischen Verortung anhand diagnostischer Strategien oder Dimensionen gar jegliche zwischen den Polen „alternativ“ und „traditionelle Leistungsmessung“. „Formative Leistungsmessungen“ können folglich mit „alternativen Formen der Leistungsmessung“ durchgeführt werden, müssen aber nicht (Maier 2014a). Das bedeutet im Umkehrschluss beispielsweise, dass eine traditionelle Klassenarbeit genauso „summativ“ oder „formativ“ genutzt werden könnte wie ein Portfolio oder eine Präsentation. Woolfolk (2012, S. 677) stellt die Zusammenhänge folgendermaßen dar:

„Die Unterscheidung zwischen formativem und summativem Erfassen leitet sich von der *Art der Verwendung* der Testergebnisse her. Jede Art der Leistungserfassung – traditionell, über gezeigte Leistung, Projekte, mündlich, Sammelmappen usw. – kann sowohl für formative als auch für summative Leistungserfassung verwendet werden.“

So kann theoretisch ein und derselbe Test zu Beginn einer Unterrichtseinheit formativ genutzt werden und am Ende der Einheit summativ (Woolfolk 2012). Noch einen Schritt weiter geht Harlen (2005, zit. nach Maier 2015, S. 41), der anführt, die Unterscheidung zwischen „formative“ und „summative assessment“ könne erst *nach* einer erfolgten Messung analytisch, also in erster Linie über die *Nutzung*, erfolgen. Klar ist jedoch, dass aufgrund der fehlenden schulrechtlichen Konsequenzen bei formativen Leistungsmessungen größere Gestaltungsspielräume bei der Form der Lernstandserhebung vorhanden sind – bzw. die Lehrkräfte diese Auffassung vertreten: Die Spannweite reicht bei „formativen Leistungsmessungen“ von einem aufwändig konstruierten Test mit mehrmaliger Durchführung und Lernverlaufsdiagnostik bis zu einer „einfachen“ formlosen Beobachtung bzw. einer simplen Frage in einem Gespräch (Vollstädt 2005; Klauer 2014; Becker 2005). Maier (2014b, S. 70) erläutert bezüglich der Formen „formativer Leistungsmessungen“, dass diese „von gezielten Schülerbeobachtungen im Unterricht über Aufgaben mit Diagnosepotenzial über Zwischentests bis hin zu standardisierten oder zentralen Testverfahren, deren Nutzung dann als formativ bezeichnet werden kann, wenn Lehrkräfte, Schülerinnen und Schüler oder auch Eltern aus den Ergebnissen Schlüsse für die Verbesserung des weiteren Lehr-Lernprozesses ziehen“, reichen. Bedacht werden sollte in diesem Kontext jedoch, dass neue Ideen und Aspekte im Bereich der Diagnostik in der Regel leichter im Feld der Modifikationsdiagnostik als in jenem der Selektionsdiagnostik umgesetzt werden können.

„Formative Leistungsmessung“ versus „Pädagogische Diagnostik“: Viele der im Zusammenhang mit einer „Neuen Prüfkultur“, „alternativen Formen der Leistungsmessung“ oder „formativen Leistungsmessungen“ dargelegten Merkmale und Konstituenten decken sich mit prinzipiellen Eigenschaften einer „Pädagogischen Diagnostik“ (Ingenkamp 1975, 1999, 1988, 1991; Ingenkamp & Lissmann 2008; Lissmann 2010; Arnold 2004). So gehört vor allem die Betonung einer *prozessorientierten* und an *Förderung* interessierten Diagnostik auch zu den Primärmerkmalen der „Pädagogischen Diagnostik“ (Ingenkamp 1975). Diese Parallelen sorgen im deutschsprachigen Raum häufig dafür, dass „Pädagogische Diagnostik“ und „formative Leistungsmessung“ – irrtümlicherweise – synonym gebraucht werden.

Unter Berücksichtigung einschlägiger Fachliteratur kann der Begriff „Pädagogische Diagnostik“ auf einer Makroebene diagnostischer Termini *zwei* sich tangierende, partiell verschränkende, jedoch nicht identische Bedeutungen einnehmen (Rheinberg 2006; Ingenkamp & Lissmann 2008):

- (1) „Pädagogische Diagnostik“ als *Überbegriff für (alle) Erkenntnisgewinne, die Grundlage für Entscheidungen im pädagogischen Kontext sind.*
- (2) „Pädagogische Diagnostik“ als *Konzept oder Programm*, welches innerhalb von (schulischen) Leistungsmessungen und (Selektions)Entscheidungen *pädagogische Komponenten* besonders betont und die Person des Lernenden ins Zentrum der Betrachtung rückt.

(1) Der Begriff „Pädagogische Diagnostik“ bezeichnet zum einen gemäß seinem Namen *alle diagnostischen Prozesse*, welche Grundlage für *pädagogische Entscheidungen* sind (Rheinberg 2006). Dies kann beispielsweise aus der häufig zitierten Definition von „Pädagogischer Diagnostik“ der Bund-Länder-Kommission (1974, S. 45) abgeleitet werden:

„Unter Pädagogischer Diagnostik werden alle Maßnahmen zur Aufhellung von Problemen und Prozessen sowie zur Messung des Lehr-Lernerfolgs und der Bildungsmöglichkeiten [...] im pädagogischen Bereich verstanden.“

Die „Pädagogische Diagnostik“, welche aus der „psychologischen“ bzw. „pädagogisch-psychologischen“ Diagnostik hervorging (Eid-Petermann 2006; Schmidt-Atzert & Amelang 2012), deckt in diesem Kontext den diagnostischen Bereich von (schulischen) Lehr-Lernprozessen ab und beinhaltet daher für diese Art bzw. Intention der Erhebungen und Entscheidungen spezifische Besonderheiten (z.B. Hesse & Latzko 2009; Topsch 2010; Sacher 2006b; Helmke 2009). So unterscheiden sich sowohl die Mittel zur Datenerhebung, als auch die Konsequenzen aus der Datenauswertung bei der „Pädagogischen Diagnostik“ deutlich von anderen, teilweise gar angrenzenden Diagnostikfeldern wie der „Psychologischen Diagnostik“ (Amelang & Zielinski 2004; Fisseni 2004). Der Begriff „Pädagogische Diagnostik“ beinhaltet somit in dieser Verwendung insgesamt, wie (Lern)Leistungen entstanden sind, welche Rahmenbedingungen für den Lernenden vorherrschten, wie Leistungen prinzipiell gemessen und beurteilt werden sollten und vor allem, welche Konsequenzen und Ziele sich aus einer Messung ergeben (Lissmann 2010; Ingenkamp & Lissmann 2008; Ingenkamp 1991).

Arnold (2004, S. 334) schränkt den Begriff dahingehend ein, dass „bildungspolitisch und wissenschaftlich initiierte Erhebungen in [...] Schulen [...] nicht als Pädagogische Diagnostik bezeichnet werden, da hier die individuelle Motivierung der Informationserhebung in der Regel fehlt.“ Mit Blick auf die Schulpraxis werden unter „Pädagogischer Diagnostik“ damit vor allem die Erhebungen verstanden, welche die Lehrkräfte schüler- oder klassenindividuell auf eine Lernsituation hin konzipieren, um „pädagogische Einzelfallentscheidungen“ zu treffen (Arnold 2004; Bund-Länder-Kommission 1974).

Maßgeblich Anteil an der Einführung bzw. terminologischen Verortung des Begriffs besitzt nach eigenen Angaben Karlheinz Ingenkamp (1975, 1988; 1991). Nach eigenen Aussagen geht der Terminus „Pädagogische Diagnostik“ auf ihn zurück und wurde Ende der 1960er Jahre „in Anlehnung an die medizinische und psychologische Diagnostik“ eingeführt (Ingenkamp 1991, S. 760). Die Nähe zur „medizinischen Diagnostik“ zeigt sich vor allem darin, dass die „Pädagogische Diagnostik“ hinsichtlich Methoden, Denkweisen und wissenschaftlich abgesicherten Verfahren auf Grundsätze der „medizinischen Diagnostik“ aufbaut (Hesse & Latzko 2009; Ingenkamp & Lissmann 2008; Leutner 2001; Fisseni 2004; Arnold 2004; Rheinberg 2006).

Ingenkamp definiert die für die Erfassung *pädagogischer Sachverhalte* spezifizierte „Pädagogische Diagnostik“ folgendermaßen (Ingenkamp & Lissmann 2008, S. 13):

„Pädagogische Diagnostik umfasst alle diagnostischen Tätigkeiten, durch die bei [...] Lernenden [...] Voraussetzungen und Bedingungen planmäßiger Lehr-Lernprozesse ermittelt, Lernprozesse analysiert und Lernergebnisse festgestellt werden, um individuelles Lernen zu optimieren.

Zur Pädagogischen Diagnostik gehören ferner die diagnostischen Tätigkeiten, die die Zuweisung zu Lerngruppen oder zu individuellen Förderprogrammen ermöglichen sowie [...] der Erteilung von Qualifikationen zum Ziel haben.

Nach der Definition Ingenkamps (Ingenkamp & Lissmann 2008) besteht folglich „Pädagogische Diagnostik“ nicht nur aus der Erhebung diagnostisch relevanter Informationen, sondern auch aus der gezielten Behandlung von Schwächen und der konkreten Angabe von potentiellen Maßnahmen (Bund-Länder-Kommission 1974). Es geht somit um mehr als eine bloße Leistungsfeststellung – übrigens analog zu allen Konzepten der „formativen Leistungsmessung“. Zum heutigen Verständnis von „Pädagogischer Diagnostik“ gehört damit auch, dass Lernergebnisse in Beziehung zu Lernvoraussetzungen gesetzt werden. Auf Basis dieser Erkenntnisse sollen weitere Maßnahmen ergriffen werden, die jedoch das gesamte Lernumfeld der Schülerinnen und Schüler betreffen (Reulecke & Rollet 1976; Helmke 2009).

(2) Über die größtenteils noch prinzipiell anerkannten Aussagen Ingenkamps (1975, 1988; 1991) bzw. Ingenkamps & Lissmanns (2008) können erste Unterschiede zwischen verschiedenen Definitionsvarianten von „Pädagogischer Diagnostik“ festgestellt werden. Diese haben zur Folge, dass der Terminus „Pädagogische Diagnostik“ nicht nur als *Überbegriff* für die allgemeine Erfassung pädagogischer Sachverhalte fungiert, sondern auch die Überlegungen *einzelner Autoren* zur Operationalisierung (schulischer) Leistungsmessung mit besonderer Betonung pädagogischer Komponenten in sich eint (Heller 1984). Da sich in diesem Falle sowohl theoretische Vorstellungen als auch praxisnahe Umsetzungsvorschläge deutlich voneinander abgrenzen können, empfiehlt es sich, in diesem Kontext nicht mehr von „Pädagogischer Diagnostik“ generell, sondern von spezifischen *Konzepten* der „Pädagogischen Diagnostik“ zu sprechen (z.B. Horstkemper 2006).

Ingenkamp (1988; 1991) selbst bilanzierte ab Mitte der 1980er Jahre, dass der Begriff „Pädagogische Diagnostik“ ungefähr zwanzig Jahre nach seiner (offiziellen) Einführung äußerst heterogen besetzt ist und somit nicht (wie erhofft) den gesamten Bereich aller Diagnostikverfahren im Bildungssektor *einheitlich* definiert. So führt er im „Lehrbuch der Pädagogischen Diagnostik“ zu diesem Sachverhalt an (Ingenkamp 1988, S. 10):

„Pädagogische Diagnostik ist auch heute noch eher ein – heftig umstrittenes und unklares – Programm als eine etablierte wissenschaftliche Teildisziplin. Daher ist es nicht verwunderlich, dass es recht unterschiedliche Definitionen gibt, was Pädagogische Diagnostik sei.“

Ein gravierender Unterschied zwischen verschiedenen *Konzepten* „Pädagogischer Diagnostik“ kann über die bereits dargelegten diagnostischen Strategien (v.a. *Selektions-* versus *Förderdiagnostik*) aufgezeigt werden. So können in den meisten Konzepten der „Pädagogischen Diagnostik“ Leistungsmessungen und -bewertungen *sowohl* zur Selektions- *als auch* zur Förderdiagnostik herangezogen werden (Arnold 2004), es existieren jedoch auch Überlegungen, bei denen der genannte Begriff eine *Abgrenzung* von selektierenden Leistungsmessungen innerhalb schulischer Lehr-Lernprozesse verdeutlichen soll (z.B. Projektgruppe 1973).

Betrachtet man den Begriff „Pädagogische Diagnostik“ im Sinne Ingenkamps (1975) als Überbegriff für alle diagnostischen Vorgänge im Feld der Pädagogik deckt der Terminus also zum einen Erhebungen zum Zwecke des *Förderns und Adaptierens* und zum anderen Messungen mit dem Ziel *Selektionsentscheidungen* durchzuführen ab (Reulecke & Rollet 1976; Klauer 1982; Rheinberg 2006; Arnold 2004). Im Gegensatz zu einer reinen Selektionsdiagnostik, wie sie in der Schule praktisch kaum vorkommt (eine Ausnahme nehmen die Schullaufbahn abschließenden Prüfungen ein), steht bei „Pädagogischer Diagnostik“ jedoch immer im Mittelpunkt, eine für den Lernenden optimale Entscheidung zu treffen (Ingenkamp & Lissmann 2008; Projektgruppe 1973). In erster Linie sollte „Pädagogisches Diagnostizieren“ folglich „solche Informationen zu Lernvoraussetzungen, Lernprozessen oder dem Lernstand eines Schülers ermitteln, die relevant sind für die gezielte Unterstützung des Schülers“ (ISB 2009, S. 5). Mit den erfassten Daten ist es dem Lehrenden möglich, folgende Lehr-Lern-Maßnahmen den individuellen Voraussetzungen des Lernenden anzupassen (Ingenkamp & Lissmann 2008; Arnold 2004). Klafki (1996) betont, dass die pädagogische Beurteilung primär Informationen *für* das Kind liefern soll. Allerdings darf die Betonung der pädagogischen Komponente einer Leistungsmessung nicht dahingehend übersimplifiziert werden, dass „Pädagogische Diagnostik“ schlicht mit einer milden Zensurierung gleichgesetzt wird – dies würde dem Sachverhalt nicht gerecht (Schröter 1981). Erhebungen im Bereich der „Pädagogische Diagnostik“ müssen folglich stets auch einem pädagogischen Maßstab gerecht werden (Weinert 2002b; Knauer 2005). Selektionsent-

scheidungen sind damit die *Folge* der angeführten Intentionen, nicht der eigentliche *Grund* der Diagnose. Daher verstehen mehrere Autoren unter „Pädagogischer Diagnostik“ eine Diagnose, welche sich durch ihre Primärintention, durch die Art ihrer Durchführung, durch ihre Auswertung und durch ihre schlussendliche Verwendung fundamental von Diagnostikverfahren anderer Bereiche unterscheidet (z.B. Amelang & Zielinski 2004; Horstkemper 2006; Bund-Länder-Kommission 1974; Ingenkamp 1991; Heller 1984).

Wie bereits in der Eingangserläuterung zu diesem Aspekt der „Pädagogischen Diagnostik“ dargestellt, existieren auch Konzepte der „Pädagogischen Diagnostik“, die keine Selektionsentscheidungen beinhalten (Projektgruppe 1973; Horstkemper 2006; Knauer 2005). In diesem Falle müsste man zwischen „schulischer Leistungsbeurteilung“ ganz allgemein und „Pädagogischer Diagnostik“ als Abgrenzung zu herkömmlicher, auf Selektion ausgerichteter Diagnostik differenzieren. Verwechslungen mit dem Begriff „formative Leistungsmessung“ rühren neben den bereits erwähnten ähnlichen Merkmalen nun eben vor allem von den (wenigen) Vertretern einer „Pädagogische Diagnostik“ her, für welche das Ziel einer derartigen Diagnose nicht über „das Erreichen oder Nichterreichen von Lernzielen und die Gründe für ein Nichterreichen“ hinausgehen darf (Projektgruppe 1973, S. 98). Soweit gehen die meisten Vertreter einer „Pädagogischer Diagnostik“ jedoch nicht, vielmehr ist es intendiert, die Schwachstellen einer oftmals praktizierten traditionellen selektiven Leistungsmessung auszugleichen (Ingenkamp 1991, 1988).

Prinzipiell kann bilanziert werden, dass die Verwendung des Zusatzes „pädagogisch“ im Zuge von Diagnostikprozessen die – im Vergleich zu vielen im Schulalltag stattfindenden „herkömmlichen“ Leistungsmessungen – stärkere Fokussierung auf den individuellen Lernfortschritt der Schülerinnen und Schüler hervorheben möchte (Hesse & Latzko 2011). Diese Zielstellung ist identisch mit der Intention der „Förderdiagnostik“ und auch bei allen Konzepten der „formativen Leistungsmessung“ prioritär. Entscheidender Unterschied ist jedoch, dass nahezu alle Konzepte der „Pädagogischen Diagnostik“ die Option für Selektions- und Modifikationsentscheidungen beinhalten (z.B. Reulecke & Rollett 1976), „formative Leistungsmessungen“ jedoch *niemals* Selektionsentscheidungen nach sich ziehen (z.B. Black & Wiliam 1998a, 1998b; Boston 2002).

„Assessment for learning“ versus „Assessment of learning“: Weiter ist zu klären, wie mit dem Begriffspaar „assessment for learning“ versus „assessment of learning“ umgegangen wird. Häufig werden beide Termini synonym für „formative assessment“ bzw. „summative assessment“ eingesetzt (z.B. Black et al. 2004; Stiggins 2002), es existieren jedoch auch andere Vorgehensweisen. Noonan & Duncan (2005) beispielsweise sehen zwar sowohl „assessment for learning“ als auch „formative assessment“ als Umsetzungen einer Reformbewegung zur Verbesserung des Lernerfolgs bei Schülerinnen und Schülern, nehmen aber darüber hinaus bei der diagnostischen Strategie der „Beobachtungsperspektive“ (Mühlig & Petermann 2006) eine Unterscheidung vor: Modifikationsdiagnostik mit Selbst- oder Peerbewertung wird als „assessment for learning“ und Modifikationsdiagnostik mit einer Bewertung durch die Lehrkraft (Fremdbewertung) als „formative assessment“ bezeichnet (Noonan & Duncan 2005). Sie begründen diese Differenzierung über die Herkunft beider Begriffe: Während „assessment for learning“ einem Konzept zur Klassenbewertung entstammt (Guskey 2003; Stiggins 2002), bezieht sich „formative assessment“ auf die Rückmeldungen durch die Lehrkraft (Noonan & Duncan 2005). Diese Aussage ist problematisch, da in vielen Konzepten zur „formativen Leistungsmessung“ „Selbsteinschätzungen“ eine bedeutsame Rolle spielen, auch der häufig zitierte Ansatz von Black & Wiliam (1998) sieht dies vor. Noonan & Duncan (2005) weisen in diesem Zusammenhang darauf hin, dass ihrem

Verständnis nach auch „formative Leistungsmessungen“ auf „Selbsteinschätzungen“ basieren können, die Lehrkraft jedoch stets am Lernprozess beteiligt ist und vor allem die Verantwortung für entsprechende Entscheidungen trifft. Schwerwiegender als die „Beobachtungsperspektive“ ist für Noonan & Duncan (2005, S. 1) somit, *wer* letztlich für „die Verbesserung des Lernenden verantwortlich ist“: Bewerten sich die Schülerinnen und Schüler selbst oder gegenseitig und entscheiden auf dieser Basis nahezu selbstverantwortlich weitere Schritte, liegt „assessment for learning“ vor, werden die Adaptionsentscheidungen von der Lehrkraft oder zumindest unter deren Beteiligung (unter anderem auch auf der Basis von Schülerselektions- und Förderdiagnostik) getroffen, sprechen Noonan & Duncan (2005) von „formative assessment“. Diese Differenzierung erscheint mit Blick auf die unterschiedlichen Einteilungskriterien diagnostischer Vorgänge auf einer rein theoretischen Ebene durchaus sinnvoll, gibt sie doch der Kombination einiger Kriterien (*Interventionsstrategie* und *Grad der Beteiligung*) einen übergeordneten Namen. Problematisch ist in diesem Zusammenhang, dass ähnlich wie beim Begriff „formative assessment“ (Synonym für „Modifikations- und Förderdiagnostik“ versus Überbegriff für eine Kombination aus „Modifikations- und Förderdiagnostik“, „Prozessdiagnostik“ sowie „während der Lernphase platzierte Diagnostik“) der Terminus „assessment for learning“ vor allem in praxisnahen bzw. keine theoretisch-differenzierende Betrachtung vornehmenden Publikationen meist synonym für „formative assessment“ verwendet wird. Somit ist oftmals unklar, ob die Formulierung „assessment for learning“ ausschließlich eine Angabe zur Interventionsstrategie („Modifikations- und Förderdiagnostik“) oder zusätzlich auch eine Aussage zur Beobachterperspektive bzw. zum Grad der Beteiligung („Selbst- oder Peerbeurteilung“) liefert. Unter Berücksichtigung der schon heterogenen Definitionssituation von „formative assessment“ soll für den weiteren Verlauf der Arbeit kein weiterer unterschiedlich zu interpretierender Terminus im gleichen Spektrum hinzukommen. Daher wird „assessment for learning“ hier „formative assessment“ gleichgesetzt, soll die Dimension der Beobachtung angeführt werden, wird diese explizit zusätzlich angeführt (Leutner 2001).

2.1.2.3 Formative Leistungsmessungen im pädagogischen Kontext

Um mit dem skizzierten Definitionsproblem von „formativen Leistungsmessungen“ umzugehen, sollen zwei Festlegungen vorgeschlagen werden: Zum einen empfiehlt es sich, diagnostische Entscheidungsfelder ausschließlich aus der *Perspektive der pädagogisch-psychologischen Diagnostik* zu bestimmen (Reulecke & Rollet 1976; Leutner 2001), und zum anderen erscheint es hilfreich, die Taxonomie von Pawlik (1976) dahingehend aufzunehmen, dass die *Intentionsstrategie* (*Selektions-* versus *Modifikationsdiagnostik*) Primärcharakter besitzt und alle weiteren Kriterien sich an dieser – teilweise unabhängig, teilweise in einem Abhängigkeitsverhältnis – ausrichten (Klauer 2014; Wholey 1996).

Greift man auf die Kategorisierungsmöglichkeit von Leutner (2001) zurück, hat man durch das Forschungsfeld des Autors bereits den Teilbereich der „pädagogisch-psychologischen Diagnostik“ ausgewählt. Leutner (2001, S. 522) differenziert diese Sparte der Diagnostik in Anlehnung an Pawliks (1976) allgemeine Aufstellung mit den zwei bekannten diagnostischen Strategien, die zuerst einmal *keine* zwangsläufige Verbindung miteinander besitzen:

- *Selektionsdiagnostik/Auslesediagnostik* vs. *Modifikationsdiagnostik/Förderdiagnostik*
- *Statusdiagnostik* vs. *Prozessdiagnostik*.

Der ausschließliche Blickwinkel auf pädagogische Diagnostikabläufe bringt bei Leutner (2001) in Anlehnung an Bloom et al. (1971) mit sich, dass die Strategie der „diagnostischen Bemühung“ durch Ersatzwörter bzw. Ergänzungen erweitert wird: Zu „Selektionsdiagnostik“

gesellt sich „Auslesediagnostik“ und zu „Modifikationsdiagnostik“ der Begriff „Förderdiagnostik“. Was auf den ersten Blick wie eine schlichte Ansammlung an Synonymen erscheint, entpuppt sich bei genauerem Hinsehen als Inhaltserweiterung: Während „Selektions- und Auslesediagnostik“ für den Bereich der Pädagogik durchaus gleichgesetzt werden können (Pawlik 1976) – selbstredend wären auch hier Abgrenzungen denkbar – sprechen Pawlik (1976) oder Scriven (1967) bei „Modifikationsdiagnostik“ noch nicht von „Förderung“. Wie bereits skizziert, versteht man bei Pawlik (1976, S. 16) unter „Modifikation“ eine „Optimierung durch eine Veränderung des Verhaltens und/oder von Bedingungen.“ Übertragen auf den schulischen Kontext würde der überwiegende Teil der Bedingungsmodifikation in den Einflussbereich der Lehrkräfte bzw. der Schule fallen (z.B. gut ausgestattete Klassenzimmer, freie Zeiteinteilung, lernförderlicher Arbeitsplatz), Verhaltensmodifikationen vor allem auf die Lernenden abzielen (neue Lernprogramme, positive Grundeinstellung zu Lernen und Schule). Gleich, welche Implementierungsrichtung (Verhaltens- versus Bedingungsmodifikation) vorliegt, spricht Pawlik (1976, S. 16) bei der Zielangabe von „Optimierung“ und „Verbesserung“, nicht von „Fördern“ – was zwar durch die Intervention des „Förderns“ erreicht werden kann, jedoch nicht deckungsgleich mit „Verbesserung“ und „Förderung“ sein muss (Gold 2011). Erst Bloom et al. (1971) setzen „Modifizieren“ mit „Fördern“ gleich und überschreiben diese Ausrichtung mit „formativem Bewerten“. Der Bezug zur Institution Schule sorgt bei der Präzisierung der diagnostischen Strategien und Dimensionen folglich dafür, dass das für Schülerfolg bedeutsame pädagogische Feld der „Förderung“ oder „Unterstützung“ (z.B. Schuck 2006) Teil der psychodiagnostischen Zielsetzungen wird. „Förderung“ stellte im Bereich der „klinischen Diagnostik“ (Pawlik 1976) bzw. der „Evaluationsdiagnostik“ (Scriven 1967) noch keine Interventionsstrategie dar.

Die Einengung der Perspektive auf pädagogische Diagnostikvorgänge bewirkt auch bei der Verortung auf der zweiten Ebene („Status- und Prozessdiagnostik“), dass im Vergleich zu Pawliks Ausführungen (1976) konkretere Aussagen anzufinden sind. So verfolgt „Statusdiagnostik“ in der pädagogisch-psychologischen Diagnostik das Ziel „den Zustand einer Person bezüglich derjenigen Aspekte des Verhaltens und Erlebens zu erfassen, die für die zu treffenden pädagogischen Entscheidungen relevant sind, wobei davon ausgegangen wird, dass dieser Zustand über einen längeren Zeitraum stabil und folglich die getroffenen Entscheidungen nur langfristig zu prüfen sind“ (Leutner 2001, S. 522). Prozessdiagnostik hingegen möchte „die Veränderungen einer Person bezüglich derjenigen Aspekte des Verhaltens und Erlebens erfassen, die für die zu treffende pädagogische Entscheidung relevant ist“ (Leutner 2001, S. 522). Dabei wird davon ausgegangen, dass sich die internen Bedingungen innerhalb eines bestimmten Zeitraums verändern können (Leutner 2001). Demnach ist Diagnose im Zuge von Lehr-Lernprozessen stets zu einem großen Teil Prozessdiagnostik, da Lernergebnisse immer als veränderbar angesehen werden müssen.

Unter Berücksichtigung der theoretischen Modelldimension (Pawlik 1976) sollten die Dimensionen als prinzipiell unabhängig voneinander behandelt werden, wobei „einige [...] Klassifikationen sinnvollerweise miteinander verbunden werden können“ (Scriven 1967, S. 152). Die weiteren Zuordnungen zu Modifikations- und Selektionsentscheidungen deklariert Scriven (1996, S. 153) als „kontextuell abhängig“ und „nicht immanent“. Wie bereits angeführt, wird in der (schulischen) Praxis Selektion jedoch meist ausschließlich über Statusdiagnostik erreicht, bei Modifikation auf der anderen Seite werden aber Methoden der Prozess- und Statusdiagnos-

tik verwendet (Pawlik 1976; Leutner 2001). Somit gibt die jeweilige Interventionsstrategie, ob als Rein- oder Mischform, die Ausrichtung der folgenden Dimensionen an. Gemäß den bisherigen Erkenntnissen liegt auf der operationalen Ebene (z.B. in schulischen Lehr- Lernprozessen) meist die folgende Zuordnung vor:

- Selektionsdiagnostik: als Statusdiagnostik am Ende eines Lehr-Lernprozesses,
- Modifikationsdiagnostik: als Prozessdiagnostik zur Förderung während eines Lehr-Lernprozesses.

Mit Blick auf die Schulpraxis wird diese Zuordnung überwiegend dann herangezogen, wenn von „summativ“ (Selektionsdiagnostik am Ende eines Lehr-Lernprozesses) bzw. „formativ“ (Förderdiagnostik während eines Lehr-Lernprozesses) gesprochen wird (Woolfolk 2012). Allerdings ist diese Bindung nicht zwangsläufig, so weist beispielsweise Scriven (1996, S. 152) daraufhin, dass „formative Evaluationen keine Form der Prozessevaluationen“ darstellen.

Weiter ist für den Bereich der pädagogischen Diagnostik bedeutsam, dass hier nicht alle potentiellen diagnostischen Zielsetzungen relevant werden. Grund hierfür ist in erster Linie das – beispielsweise im Vergleich zur klinischen Psychologie – stark eingeschränkte Messspektrum. Greift man auf die nun bereits häufig zitierte Standardeinteilung für diagnostische Prozesse von Pawlik (1976, S. 32) zurück (Abb. 1),

- so liegt bei pädagogischen Maßnahmen in der zweiten „Alternativdimension psychodiagnostischer Zielsetzungen“ nahezu immer „*kriterienorientierte*“ und nicht „*normorientierte*“ Diagnostik vor;
- so ist die dritte „Alternativdimension“ („*Testen* versus *Inventarisieren*“) mit wenigen Ausnahmen nur „im klinisch-psychologischen Bereich und hier insbesondere in der Verhaltenstherapie“ bedeutsam;
- und so sind mit Ausnahme von Abschlussprüfungen im schulischen Kontext fast ausnahmslos Diagnostikvorgänge als „Information“ und nicht nur als „Messung“ anzufinden, da sogar bei Selektionsentscheidungen „das Einbringen von Informationen für und über Behandlung [Lehr-Lernprozesse]“ als Zielsetzung angenommen wird (S. 34).

Für die Schulpraxis bedeutet diese Verortung, dass hier im Gegensatz zum klinisch-psychologischen Diagnostikspektrum oftmals *keine* Wahlmöglichkeit hinsichtlich der Strategien bzw. Dimensionen besteht. Diese Situation liegt partiell sogar dann noch vor, wenn man sich den „diagnostischen Entscheidungsfeldern“ der „pädagogisch-psychologischen Diagnostik“ von Leutner (2001) widmet, welche schon einen Fokus auf pädagogische Belange richten und die Auswahl Pawliks (1976) stark einschränken. Neben den auch tatsächlich von den Lehrkräften festzulegenden Strategien (*Selektions-* versus *Förderdiagnostik* und *Status-* versus *Prozessdiagnostik*) existieren somit weitere Strategien, die zwar zur Kategorisierung genutzt, oftmals jedoch von der Lehrkraft nicht aktiv gewählt werden können.

Unter Berücksichtigung der Aufstellungen von Leutner (2001) und Pawlik (1976) kann für die Schulrealität folgende Auswahl an Dimensionen angeführt werden, die (v.a. durch die Lehrkraft) tatsächlich eine Manipulation erfahren (Abb. 2). Neben den bereits mehrfach erwähnten Dimensionen „*Intention*“ und „*Zustand*“ der Leistungsmessung“ können schulische Leistungsmessungen vor allem über den „*Zeitpunkt*“, den „*Gegenstand*“, die „*Perspektive*“, die Form sowie die „*Bezugsnorm*“ beschrieben werden (Abb. 2).

Darstellung schulischer Diagnostikprozesse

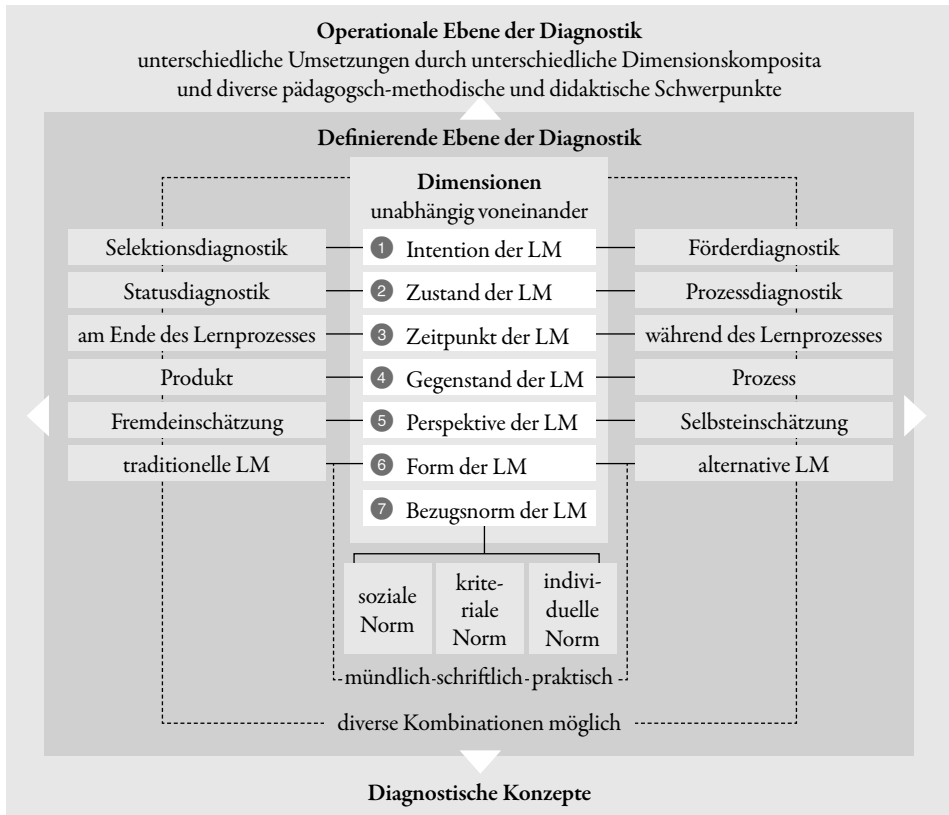


Abb. 2: Auswahl diagnostischer Strategien bzw. Dimensionen

Vor allem im Kontext schulischer Lehr- Lernprozesse wird in der Regel die Dimension „*Intention*“ und gleichzeitig die Dimension „*Zustand*“ belegt, wenn von „formativen“ oder „summativen“ Leistungsmessungen gesprochen wird (z.B. Woolfolk 2012). Wie über die theoretischen Strategien Pawliks (1976) bereits dargestellt, wird partiell auch im pädagogischen Kontext nur die Dimension „*Intention*“ als Kriterium zur Manifestierung der Termini „formativ“ und „summativ“ herangezogen, dies ist zum einen aber eher selten der Fall und zum anderen erhöht diese Perspektive die skizzierten Definitionsschwierigkeiten erheblich (Scriven 1996). Um die angeführte Problematik zu entschärfen und beide Perspektiven (*Beschreibung anhand der Intention und des Zustandes* vs. *Beschreibung anhand ausschließlich der Intention*) zu einen, wird im deutschsprachigen Bereich oftmals von einer „*Prozessorientierung*“ (statt *Prozessdiagnostik*) gesprochen, was den Schwerpunkt auf prozessdiagnostische Vorgänge, sprich mehrfache Messungen eines *veränderbaren* Konstrukts *während* eines Lehr-Lernprozesses, betont, einzelne Messungen – unter Umständen auch am Ende einer Lehrereinheit – jedoch nicht ausschließt. Vollstädt (2005) weist in diesem Zusammenhang beispielsweise daraufhin, dass in der Schulpraxis durchaus Konzepte „*summativer Leistungsmessungen*“ Anwendung finden, die nicht am (vorläufigen) Ende eines Lernprozesses stattfinden und „*formativ genutzte Messungen*“ durch-

geführt werden, welche nicht während eines Lernprozesses platziert sind. Hier bezeichnen die Begriffe „formativ“ und „summativ“ somit ausschließlich die Intentionstrategie (auch Henry et al. 2013). Dies kann auf der operationalen Ebene beispielsweise bedeuten, dass ein Lernprozess (vorläufig) abgeschlossen wird, ohne dass überhaupt eine selektive Leistungsbewertung stattfindet. Um die Schülerinnen und Schüler aber dennoch über ihren Leistungsstand zu informieren, kann auf eine formative Leistungsmessung zurückgegriffen werden, welche dann am Ende einer längeren Lernphase steht. Vor allem am Ende von Projekten ist eine Rückmeldung zur Projektphase und einem möglichen Ergebnis üblich und sinnvoll, diese muss aber nicht über Zensuren erfolgen (Bastian 1997).

Als Fazit für den *Kontext von Schule* kann folglich herausgestellt werden, dass die *intentionale* Entscheidung in einem Diagnostikprozess in jedem Falle Primärcharakter besitzt. Damit eint alle Ansätze formativ genutzte Leistungserhebungen die *Intention*, Lernen zu unterstützen bzw. zu fördern. Maier (2010 S. 293) spricht in diesem Zusammenhang noch allgemeiner von einer „Optimierung von Lehr-Lernprozessen“, wenn es um die Bestimmung des Hauptanliegens formativer Leistungsmessungen geht. Wie bereits erwähnt, findet man diese Grundintention demnach auch in nahezu allen gängigen Erläuterungen von „formative assessment“ (z.B. Black & Wiliam 2009; Prengel et al. 2009; Marsh 2004; Yorke 2003; Swearingen 2002; Sutton 1992; Sadler 1989). Weiter kann für schulische Lehr-Lernprozesse bilanziert werden, dass die *Intention* der Leistungsmessung nicht zwingend den *Zeitpunkt* und die *Form* dieser Messung vorgibt, sie jedoch wie bei weiteren Zieldimensionen (Abb. 2) ausrichtend fungiert (Pawlik 1976). In der Schulpraxis beeinflusst das *Ziel* einer Lernstandserhebung meist *Form* und *Planung* (also auch den Zeitpunkt) dieser massiv (Zhang 2005): Versuchen die Lehrkräfte bei „summativen Leistungsmessungen“ noch immer in erster Linie schulrechtliche Vorgaben über bewährte und traditionelle Messinstrumente zu berücksichtigen und abgeschlossene Lehr- Lernprozesse zu prüfen (Paradies et al. 2009; Becker 2005), steht bei „formativen Leistungsmessungen“ die Aussagekraft und die Qualität der Rückmeldung über einen individuellen Lernstand bzw. dessen Entwicklung noch während des Entwicklungsprozesses im Vordergrund (Gold 2011; Klauer 2014; Becker 2005; Maier 2014a). Das Instrument einer „formativen Leistungsmessung“ ist somit oftmals darauf ausgerichtet, schon kleinste Lernfortschritte zu erfassen, was in der Regel motivierend auf den Lerner wirkt (Vollstädt 2005; Klauer 2014). Inhaltlicher Bezugspunkt der Förderdiagnostik ist mit Blick auf die häufig in einen Lernprozess integrierte Messung die sich Schritt für Schritt auszubildende Gesamtkompetenz, deren augenblickliche Ausprägung über eine Lernstandserfassung erkannt wird. Ein Instrument muss daher in der Lage sein, Förderpotenzial prinzipiell aufzudecken (Gold 2011). Das bedeutet, es muss Lernstände, die noch vor den augenblicklichen Unterrichtsinhalten liegen, und Kompetenzausprägungen, welche in der Schule erst zu einem späteren Zeitpunkt erwartet werden, erfassen können (Peschel 2006b) – so wird durch die „formative Leistungsmessung“ Über- und Unterforderung vermieden. Das ermittelte Förderpotenzial soll das Einleiten von geeigneten Fördermaßnahmen, die aus „individuellen Trainingsmaßnahmen“ und „adaptiven unterrichtlichen Maßnahmen“ bestehen, ermöglichen (Gold 2011, S. 223). Die angeführten Aspekte haben in der Praxis zur Folge, dass das bestehende Instrumentarium zur Ermittlung und Bewertung schulischer Leistungen eine deutliche Erweiterung erfahren hat und stets weiter erfährt (Vollstädt 2005). Konsequenz dieser Entwicklung ist die bereits mehrfach erwähnte Heterogenität hinter dem übergreifenden Begriff „formative assessment“, welche nicht nur unterschiedliche Instrumente, sondern völlig unterschiedliche Konzepte und Schwerpunkte beinhaltet.